PREPARING WHOLE GENOME HUMAN MITOCHONDRIAL DNA LIBRARIES FOR NEXT GENERATION SEQUENCING USING ILLUMINA® NEXTERA® XT

By

Hilde Stawski

A Thesis Submitted to the Faculty of the Graduate School
of Western Carolina University in Partial Fulfillment of
the Requirements for the Degree of Master of Science

Director:

Dr. Mark R. Wilson

Associate Professor, Department of Chemistry and Physics

Director, Forensic Science Program

Committee Members:

Dr. Indrani Bose, Department of Biology

Dr. Patricia Foley, Department of Chemistry and Physics

December 2013

ACKNOWLEDGEMENTS

First of all, I would like to thank Dr. Mark Wilson, for the opportunities that were extended to me in these past years. Erin Burnside and Brittania Bintz also deserve a (metric) ton of thanks for all of their help during my time at Western Carolina University. I owe gratitude to Dr. Seán O'Connell for being my last-minute reader, and to my committee members, Dr. Indrani Bose and Dr. Patricia Foley, who guided me both professionally and personally.

In addition, thanks go to the National Institute of Justice, through which part of my work was funded, and to Illumina®, especially Drs. Cydne Holt, Kathryn Stephens and Joe Varlaro, for providing us with both materials and knowledge.

Most importantly, I would like to thank my family and friends, both in the United States and at home in the Netherlands. Without their support I would certainly have lost the remainder of my sanity. In particular, thanks go to my mother, for putting up with my absence, my brother, for his never-ending blunt humor, and Marco, for being in my life. Lastly, I wish to thank my father, who was the strongest man I have ever known.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

PREPARING WHOLE GENOME HUMAN MITOCHONDRIAL DNA LIBRARIES FOR NEXT GENERATION SEQUENCING USING ILLUMINA® NEXTERA® XT

Hilde Stawski, B.S.

Western Carolina University (December 2013)

Director: Dr. Mark R. Wilson

Forensic DNA casework principally relies on the analysis of short tandem repeats (STRs) from nuclear DNA (nDNA). In cases where nDNA may not be suitable for analysis (i.e., highly degraded DNA or DNA present in quantities too low to obtain an STR profile), mitochondrial DNA (mtDNA) is an excellent alternative. MtDNA is a circular genome of approximately 16.5 kb, is maternally derived, and is present in thousands of copies per cell versus two copies of nuclear DNA. The combined higher copy number, circular shape of the genome and protection by the double membrane of the mitochondrion allows for a greater probability to recover sufficient mtDNA for typing of degraded samples.

Presently, forensic analysts sequence two or three hypervariable (HV) regions found in the non-coding control region of the mtGenome, since sequencing of the entire mitochondrial genome (mtGenome) is rather costly and labor-intensive. Additionally, difficulties sequencing through homopolymeric regions, as well as the presence of low-level mixtures in samples, can add complexity to the analysis of mtDNA in casework when traditional Sanger sequencing methods are used. These issues can be addressed

with Next Generation Sequencing (NGS) technologies. NGS enables deeper analysis of the genome for identification of low-level mixtures, since clonal populations of molecules originating from a single template strand are sequenced. Moreover, this technology allows for the more cost-effective sequencing of whole mtGenomes compared to Sanger methods, since more sequences are obtained for the same sample. By expanding mtDNA analysis to the entire mtGenome, a better resolution in distinguishing between haplotypes is established.

In forensic casework, amplification of challenging samples such as hair and aged bone is often performed differently than that of reference samples (buccal swabs, blood, etc.) due to the higher possibility of DNA degradation and limited mtDNA concentrations. For this study, two sample preparation approaches were developed including one method for robust reference samples, and one method for forensically relevant challenging samples.

For NGS analysis of reference samples, DNA was extracted from buccal swabs obtained from eight donors. A long PCR approach, which refers to the amplification of DNA fragments of a size that may not be amplified using conventional PCR reagents, was successfully performed on these DNA extracts using a highly processive polymerase mixture and novel primer pairs to amplify the mtGenome in two independent PCR reactions, with overlap at the noncoding region. These samples were subsequently processed with Illumina® Nextera® XT. This NGS library preparation kit is designed exclusively for use with Illumina® instrumentation and employs an engineered Transposome™ to randomly fragment and tag amplicons and small genomes with Illumina® specific adapters. After library preparation, samples were sequenced on the

Illumina® MiSeq™. This method generated whole mtGenome NGS data, which accurately reflected the Sanger sequence.

For analysis of challenging samples, DNA was extracted from 2 cm fragments of hair shafts from a subset of the same donors, using an optimized DNA extraction protocol. Whole genome amplification (WGA) was performed on these extracts with four different commercially available WGA kits. WGA allows for pre-amplification of the entire mtGenome without the need for any additional primer design, after which the resulting DNA can be used for downstream applications. This potentially provides the forensic analyst with an increase in DNA template, resulting in a higher possibility of obtaining useful data from a casework sample. The increase in mtDNA copy number was assessed with a human mtDNA specific qPCR assay. A subset of the samples before and after WGA was amplified using a targeted multiplex PCR approach. This product, in addition to a subset of WGA product that was not PCR amplified after WGA, was prepared with Illumina® Nextera® XT and sequenced on the Illumina® MiSeq™.

This research effort generated a protocol for obtaining whole mtGenome NGS data from reference samples such as buccal swabs. In addition, preliminary data was generated for future studies designed to obtain whole mtGenome NGS data from challenging sample types.

CHAPTER 1: INTRODUCTION

1.1 Forensic Human Identification

When a crime has been committed, samples are obtained from the crime scene and sent to the forensic laboratory. In a typical case, a Known (K) sample is taken from an individual, usually in the form of a buccal swab, and the results from the DNA analysis are compared to the Questioned (Q) samples found at the crime scene. Afterward, a statement is made as to whether or not the donor of the K sample could have contributed the Q sample.

When possible, analysis is performed on nuclear DNA (nDNA). This 3.3 billion base pairs (bp) long chromosomal DNA resides in the nucleus and is present in two copies per cell. It recombines during meiosis before it is passed down from both parents to their children. In the early stages of human identification, Dr. Alec Jeffreys discovered long repeating sequences in the nDNA and investigated these by performing Restriction Fragment Length Polymorphism (RFLP) studies. This technique employs enzymes that cut the DNA at designated sites, which shows a particular pattern upon size separation by gel electrophoresis (Jeffreys, Wilson, and Thein 1985).

Currently, forensic DNA casework largely relies on the analysis of short tandem repeats (STRs) from nDNA. These short DNA sequence repeats occur in abundance throughout the non-coding regions of the nuclear genome (Kimpton et al. 1993). In casework, the STRs are amplified with primers that are appended with fluorescent dyes, and the resulting amplicons are separated by size with capillary electrophoresis to

determine the variation in repeats between individuals (Collins et al. 2004). Since nDNA recombines in a loci-independent manner, it is very suitable for human identification because this allows for combining the results for all STR loci. By doing so it is possible to establish identity from a DNA sample.

In some cases, STRs may not be suitable for analysis (i.e. highly degraded DNA, or DNA present in quantities too low to obtain an STR profile). This is often the case with hairs and aged teeth and bone samples. In these instances, mitochondrial DNA (mtDNA) is an excellent alternative method of DNA analysis. Unlike nDNA, mtDNA is maternally inherited, which limits its resolution to a maternal lineage. However, this feature makes it particularly useful in kinship testing and the identification of human remains.

The mitochondrial genome (mtGenome) resides in the mitochondria of a cell. It is circular, approximately 16.5 kb long and is present in thousands of copies depending on the cell type, compared to the two copies of nDNA in a single cell (Budowle et al. 2003, Bogenhagen and Clayton 1974). Because of its smaller size, the total amount of mtDNA in a cell is lower than that of the nDNA, but its higher copy number, its shape and the extra level of protection from the environment provided by the double membrane of the mitochondrion increases the level of sensitivity and allows for a greater probability of recovering sufficient mtDNA for typing of degraded samples (Foran 2006).

Due to its maternal inheritance, the evidentiary value of mtDNA differs from that of nDNA; it cannot be used to establish identity. Since there is no recombination of the mtDNA, all maternal relatives have the same profile, apart from germ-line mutations that may have arisen.

When mtDNA sequence data is obtained from a questioned sample, it is compared to a reference sequence, currently the revised Cambridge Reference Sequence (rCRS) (Anderson et al. 1981), and variants from the reference are reported. The variants obtained for the K sample are then compared to those of the Q sample to determine whether or not the K sample could have contributed to the mtDNA profile of the Q sample (Figure 1) and are scored as follows (Budowle, Wilson and DiZinno, 1999):

1. "Cannot be excluded" if the Q sample and K sample have a common base at each position. This includes low-level mixed positions, which present as a mixture of nucleotides at the same position that arise due to mutations.
2. "Inconclusive" if there is one single-nucleotide polymorphism (SNP) difference between the Q sequence and the K sequence.
3. "Excluded" if there are two or more SNP differences.

If the known sample cannot be excluded as a potential source of the questioned sample, a statistical interpretation of the frequency of its haplotype in an mtDNA population database is given to provide a weight assessment for the DNA evidence. A population database is a collection of mtDNA sequences that have been observed in the population, which can be clustered in haplogroups depending on their patterns of common ancestry (Budowle et al. 2003). For example, for an "H1a1b" haplotype the "1a1b" designation refers to a particular sub-type identified within the "H" haplogroup.

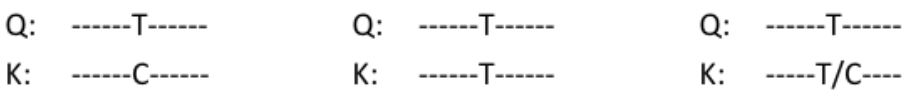


Figure 1: Different interpretations in mtDNA casework. Left: one SNP difference between Q and K, inconclusive. Middle: no differences, cannot be excluded. Right: common variant from the rCRS shared by Q and K, cannot be excluded.

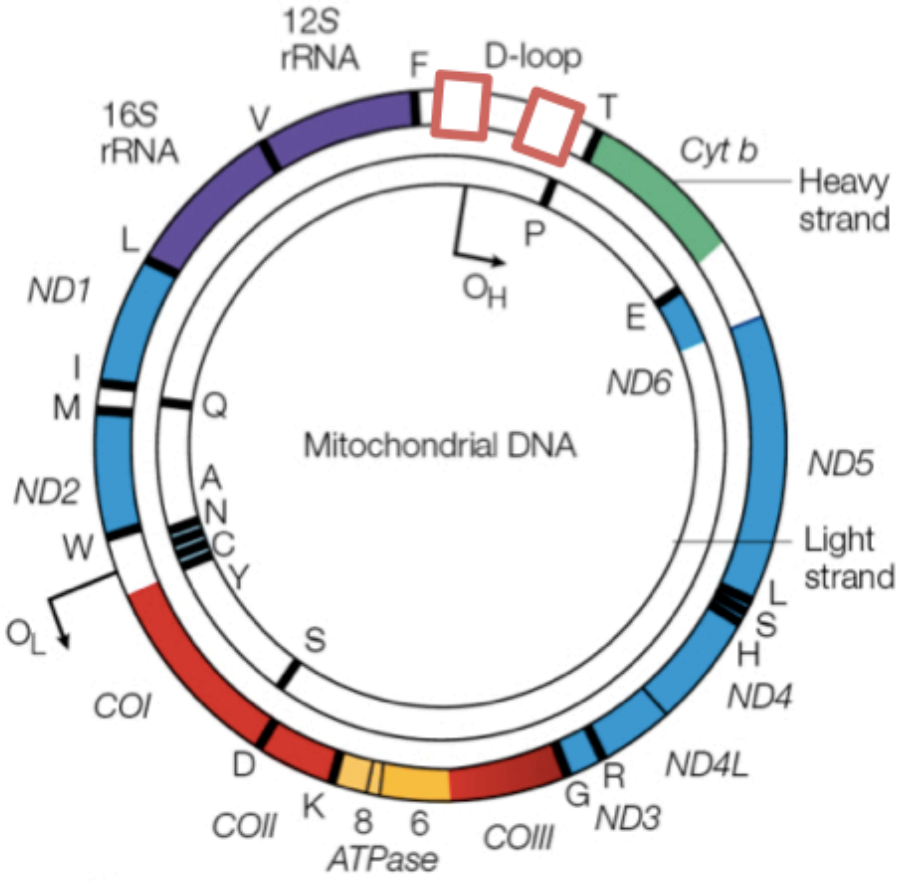


Figure 2: Structure of the mitochondrial genome (Taylor and Turnbull 2005). The hypervariable regions make up less than 4% of the entire mtGenome. Red boxes: HVI (left), HVII (right).

The mitochondrial genome itself is categorized into a coding and a non-coding region. As the name implies, the coding region codes for certain genes (Figure 2). The non-coding control region (D-loop) of the mtGenome mutates at a higher rate than the coding region (Greenberg, Newbold, and Sugino 1983, Parsons et al. 1997). This allows forensic DNA analysts to focus their analysis on two or three hypervariable (HV) regions found in the D-loop, which contain the most inter-individual variation.

Sequencing of the entire mtGenome with traditional sequencing methods (detailed in section 1.1.1) is labor intensive, and can require large amounts of DNA

(Holland MM, Parsons TJ 1999, de Vries et al. 2012). Although it has been shown to allow for a better resolution in distinguishing between haplotypes (Andréasson et al. 2007) this type of analysis is typically not performed. A higher degree of resolution is desirable in mtDNA casework, as this provides the DNA analyst with more data to provide enhanced resolution of haplotypes. The introduction of Next Generation Sequencing (NGS) instruments, which combine many sequencing reactions at once and allow for more data to be obtained from DNA samples, renders whole mtGenome sequence analysis more feasible. NGS is explained in detail in section 1.1.2.

In addition, NGS methods may assist in the identification of DNA mixtures and/or heteroplasmy, which is the existence of two or more mtDNA sequences in one individual (Salas, Lareu, and Carracedo 2001). Both length heteroplasmy - different lengths of a sequence motif - as well as sequence heteroplasmy, which manifests as more than one base at the same position, occur in mtDNA (Budowle et al. 2003, Salas, Lareu, and Carracedo 2001). The degree to which sequence heteroplasmy can be identified with traditional sequencing methods is limited (Holland MM, Parsons TJ 1999). The estimated limit of detection using currently employed methods is approximately 10% (Wilson et al. 1995).

1.1.1 Sanger Sequencing

Traditionally, mtDNA sequence analysis is accomplished using PCR followed by Sanger sequencing (Sanger, Nicklen, and Coulson 1977). With this technique the amplified DNA template is terminated at different base positions, which ultimately allows for the visualization of each base position in a sequence.

In one application of Sanger sequencing, a single forward or reverse primer is annealed to the template and extended with deoxyribonucleotide triphosphates (dNTPs) and four individual dideoxyribonucleoside triphosphates (ddNTPs) that are fluorescently labeled with four different fluorophores. A ddNTP terminates the extension of a DNA strand because it lacks a hydroxyl group on the 3' carbon to which the next dNTP is added. Termination occurs randomly at all base positions, resulting in the creation of fluorescently labeled fragments of different lengths (Figure 3). This technique is currently combined with capillary electrophoresis-based separation of the labeled fragments (Sanger, Nicklen, and Coulson 1977).

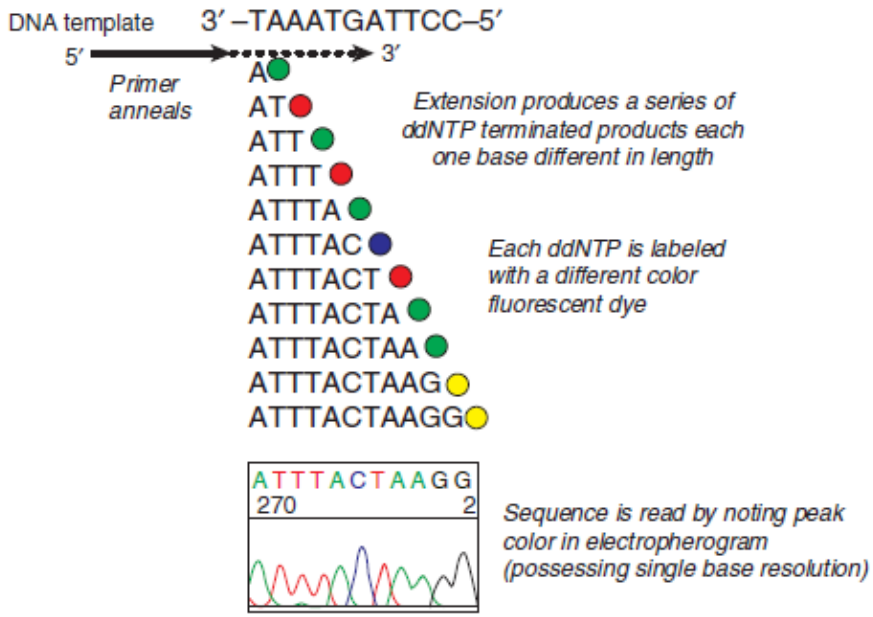


Figure 3: Sanger sequencing results in the formation of fluorescently labeled fragments of different lengths (Butler 2005).

1.1.2 Next Generation Sequencing

NGS, also called massively parallel sequencing, is a technique that combines hundreds of thousands, or on some platforms even millions, of individual sequencing reactions simultaneously (Metzker 2010, Shendure and Ji 2008). Large volumes of data are obtained by imaging the sequencing reactions in real-time.

By introducing NGS technology into the crime laboratory, more information can be obtained from the same DNA sample. With Sanger sequencing, a sequence is usually covered by one forward and one reverse "read". A read count, or coverage, defines how many times a DNA strand has been sequenced. With NGS technologies, thousands of independent reads overlap at each position, and therefore more data points are obtained from the same position. Thus, bases are sequenced at a higher depth, which increases the ability to detect heteroplasmy and/or mixtures (Figure 4). Previous studies have shown that mixtures occurring at a rate as low as 1.3-1.6% can be reliably detected with NGS, which is significantly lower than the aforementioned 10% with Sanger sequencing (Yiping He et al. 2010, Tang et al. 2013).

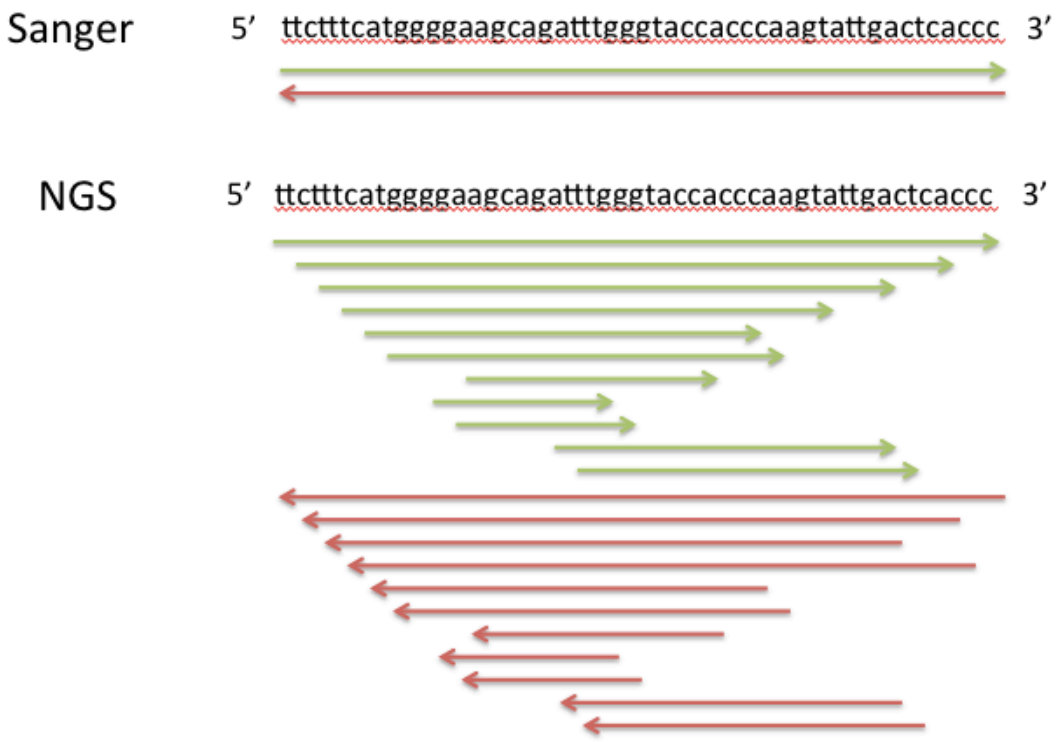


Figure 4: Differences in read depth in Sanger sequencing vs. NGS. Green: forward reads. Red: reverse reads. Additional data points are obtained using NGS due to the parallel processing of many more reads.

NGS significantly reduces the cost of sequencing (Wetterstrand KA). Although the initial instrument purchase and reagents can be quite expensive, they are used more efficiently (Liu et al. 2012, Loman et al. 2012). By incorporating short oligos of known sequence into the DNA molecules of each sample (indexing) the sequence data from each sample can be recognized during data analysis. Thus, individual sequence data from many samples can be processed simultaneously (multiplexing). This reduces the overall cost per sample, although the number of samples that can be processed is dependent on how many indices can be used in the same run. As stated previously, many more reads can be obtained for a DNA sample with NGS, which not only facilitates the sequencing of DNA in a higher degree of depth, but also breadth: more sequencing data can be

obtained across the genome. These two factors make the expansion of analysis to the entire mtGenome much more feasible on NGS systems as compared to Sanger sequencing.

There are several other differences between current NGS technologies and Sanger sequencing: sequence chemistry, read length, sequence time and sequence accuracy (Loman et al. 2012, Jünemann et al. 2013). The following bench top NGS platforms are popular in the scientific community, but many other systems also exist. The Roche® 454 GS Junior™, which measures light emission after base incorporation, can sequence up to 500 consecutive bases. The Ion Torrent™ PGM™ measures fluctuations in pH that are influenced by base incorporation, and can reach read lengths of 400 bp. The Illumina® MiSeq™ is based on the incorporation of fluorescent bases (further explained in section 1.3). Its current read length is 2x300 base pairs (300 bases forward and reverse). These maximum read lengths are shorter than that of traditional Sanger sequencing, which can generate up to 1000 bases of sequencing data (Shendure and Ji 2008). However, as stated before, NGS allows for sequencing at a higher breadth and depth than Sanger sequencing since more data points are obtained in one run.

In addition to chemistry and read length, NGS differs from Sanger sequencing in run time. Sequencing can take from several hours to several days, depending on the platform (Loman et al. 2012). Although this is significantly longer than Sanger sequencing runs, which typically last no more than a few hours (Liu et al. 2012) it is important to again note that one NGS run can result in many more data points than a Sanger run.

The raw base accuracy of the NGS instruments can be lower than capillary

electrophoresis instruments. For example, the Roche® 454 GS Junior™ shows higher rates of substitutions, insertions and deletions when sequencing through homopolymeric stretches (Loman et al. 2012). When identical nucleotides are incorporated simultaneously, the emitted light signal does not linearly increase with the number of consecutive bases in these stretches, which can cause ambiguity in length determination. However, because NGS technology provides deeper coverage and since these systems employ algorithms for error correction, these issues can be partially circumvented (Jeck et al. 2007, Kao, Chan, and Song 2011, Meacham et al. 2011).

1.2 Illumina® MiSeq™

This study focuses on the Illumina® MiSeq™ instrument. The general workflow for analysis on the MiSeq™ includes library preparation, single molecule amplification by bridge PCR and reversible terminator sequencing-by-synthesis (Shendure and Ji 2008).

Libraries are DNA fragments from the target sequence that are appended with adapter sequences. These adapters are necessary so that the target DNA will bind to complementary oligonucleotides that are bound to the surface of the reaction chamber, which is where the sequencing chemistry takes place; on the Illumina® MiSeq™ this surface is called a flow cell. This MiSeq™ flow cell is an optically transparent reaction chamber, which allows for the detection of fluorescence emanating from bound DNA sequences. The process of library preparation is detailed in section 1.3. Although other library preparations are available, this study focuses on Illumina® Nextera® XT, a

method that is commercially available through Illumina®.

After binding of the libraries to the flow cell, each anchored oligo is extended; the incorporated bases are complementary to the bound DNA template (Figure 5A). This double stranded DNA is denatured and the original strand is washed away, leaving only the covalently anchored libraries. Bridge PCR (bPCR) is then used to clonally amplify each single bound DNA molecule (Bentley et al. 2008).

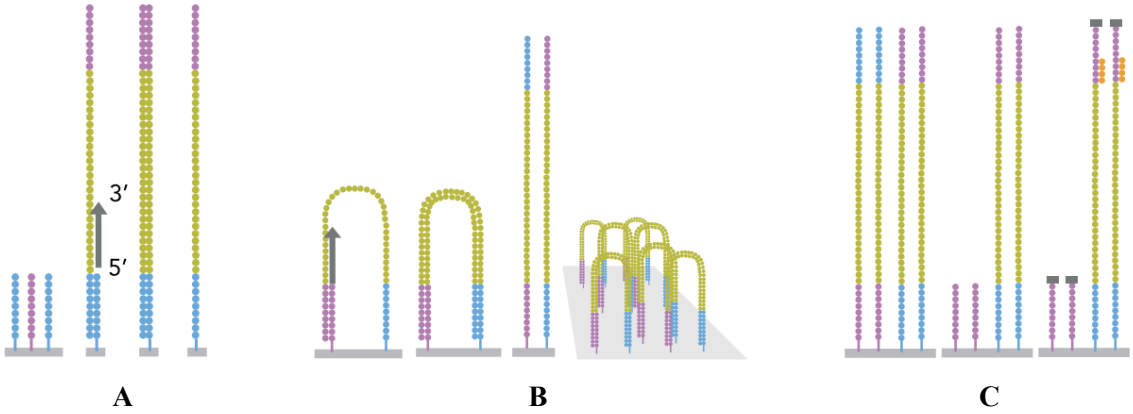


Figure 5: Bridge PCR and preparation of libraries for Illumina® sequencing (Illumina® 2011). A: Libraries attach and are immobilized after which the strands are extended in the 3' direction. B: Bridge PCR is performed, creating clusters of each single DNA molecule. C: Strands are linearized and all 3' ends are blocked prior to sequencing.

The opposing end of each synthesized strand hybridizes to complementary lawn primers, forming a bridge (Figure 5B). Primers are added and extended, after which the resulting DNA strands are denatured. Hybridization and extension steps are repeated throughout multiple cycles, generating a so-called "cluster" of approximately a thousand copies from a single DNA molecule. All fragments of the original template are amplified and hybridized in this way, resulting in billions of individual sequences generated in parallel (Shendure and Ji 2008, Metzker 2010).

Bridge PCR is followed by reversible terminator sequencing-by-synthesis. This

process starts with denaturation of the clusters, after which the reverse strands are cleaved off to leave ssDNA fragments bound to the flow cell (Figure 5C). Then, the 3' ends of the remaining forward strands and the unused flow cell-bound oligos are blocked with ddNTPs by a terminal transferase. Sequencing primers are then added, and hybridized to the immobilized fragments. These sequencing primers are extended by one base with fluorescently labeled dNTPs, which are complementary to the template bound to the flow cell. These dNTPs are blocked with a 3′-O-azidomethyl blocking group (Figure 6A) (Bentley et al. 2008). The labels are excited with a laser and emit light signals at defined wavelengths depending upon on the incorporated, complementary, nucleotide.

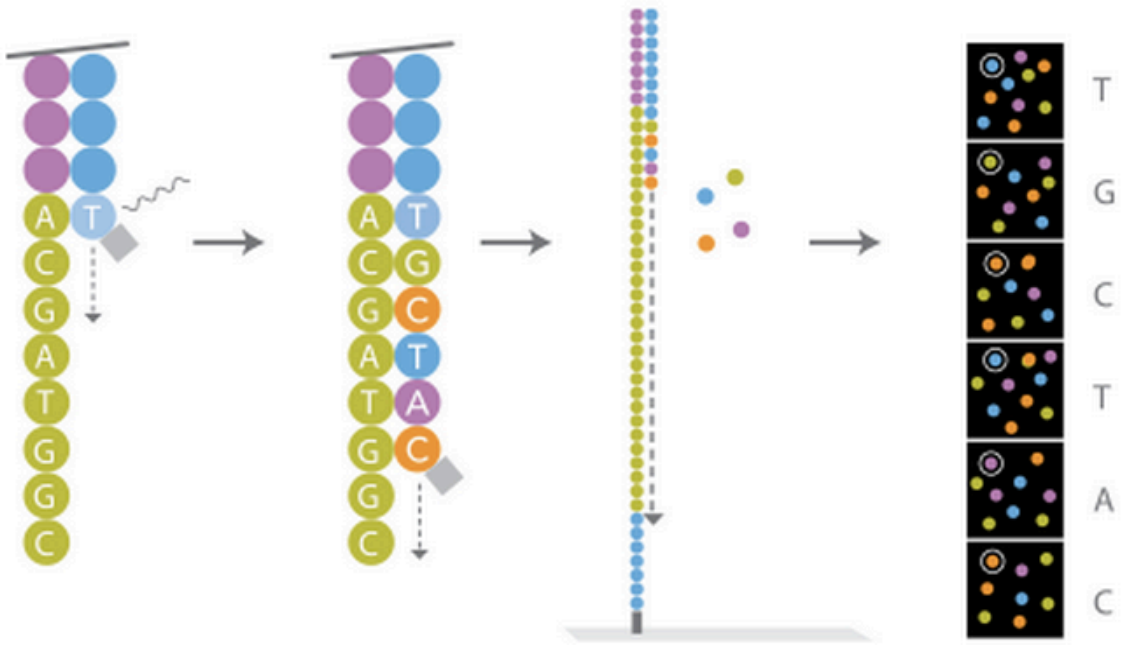


Figure 6: Reversible terminator sequencing-by-synthesis on Illumina® systems (Illumina® 2010). A: DNA template is extended with fluorescently labeled nucleotides, one nucleotide position at a time. B: Fluorescent signal from clusters is imaged by a CCD camera in real-time.

The next sequencing cycle starts with cleavage of the fluorophore and blocking group by tris(2-carboxyethyl)phosphine (TCEP) from the dNTP, followed by a wash to

remove used reagents, after which the template DNA is extended by the next dNTP (Mardis 2008). The fluorescence from all clusters on the flow cell is measured by a charge-coupled device (CCD) (Figure 6B).

The Illumina® MiSeq™ instrument features paired-end sequencing. After sequencing in the forward direction is complete, one round of bridge amplification is then performed, and the opposing end of the DNA fragments are hybridized to the flow cell. This allows for the reverse strand of the DNA fragment to then be sequenced, resulting in a doubling of the length that can be sequenced from each cluster (Figure 7) (Glenn 2011).

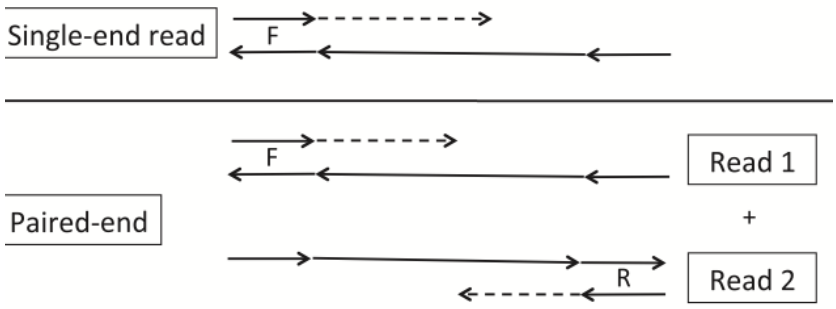


Figure 7: Single-end vs. paired end sequencing (Glenn 2011). F indicates forward read, R indicates reverse read. A single-end run is only comprised of one read, whereas a paired-end run sequences the DNA molecule from both strands in separate, linked, reads.

1.3 Illumina® Nextera® XT Library Preparation

As stated previously, the initial step in the NGS workflow is library preparation. The library consists of DNA target molecules with adapters, such that they are able to hybridize to the reaction chamber of the NGS instrument, the flow cell in the case of the Illumina® MiSeq™.

In one application, targeted amplification with modified primers: these primers include adapter sequences at the 5' ends, and the resulting amplicons can be directly

hybridized. Other strategies for library preparation have also been developed. Among these is the fragmentation of long strands of template DNA prior to addition (ligation) of the adapters.

There are several methods of fragmenting DNA molecules to prepare libraries for NGS. Fragmentation of DNA occurs by focused acoustics, nebulization or enzymatic digestion. Focused acoustics shears DNA by implementing acoustic wave energy, which results in the formation and collapse of air bubbles that causes the breakage of DNA strands (Voelkerding, Dames, and Durtschi 2010). These types of instruments require a significant starting investment, and are therefore less feasible for smaller laboratories. Nebulization is a more inexpensive technique, which harnesses the force of compressed air to shear DNA. However, this method is more prone to sample loss and cross-contamination (Voelkerding, Dames, and Durtschi 2010, Liu 2011).

Another less costly and less contamination-prone alternative to library preparation is enzymatic digestion. Different NGS platforms require different fragment sizes, and hence an advantage of enzymatic fragmentation techniques over nebulization or sonication is that the fragment size can be more easily controlled. One example includes NEBNext® dsDNA Fragmentase®, a dual enzyme system. The first enzyme randomly nicks the DNA, the other enzyme then recognizes the nick and cuts the opposite side of the strand, resulting in random fragmentation (Liu 2011, Knierim et al. 2011).

Fragmentation using some of the techniques mentioned previously result in terminal overhangs on the DNA fragments. These require repair by end blunting, or filling the terminal overhang with its complementary base. This, depending on the NGS

platform, may then be followed by addition of a single adenine base to the 3' ends (monoadenylation) so the adapters, which have a thymine overhang, can be ligated to the DNA fragment (Voelkerding, Dames, and Durtschi 2010, Liu 2011).

These processes of DNA fragmentation, end repair and adapter ligation can be very time consuming. By contrast, Illumina® Nextera® XT is a library preparation method that may potentially simplify the library preparation process. Nextera® XT has been developed for library preparation of small genomes and PCR amplicons (Illumina® 2012). It is based upon the Nextera® system that can be used to prepare larger genomes for sequencing.

Nextera® XT employs a Transposome™ complex, consisting of a transposon with free ends and a transposase enzyme (Marine et al. 2011). Transposons are genetic elements capable of cutting and pasting themselves into different locations in the genome (Campbell and Reece 2004). They typically contain a gene that codes for the transposase enzyme, which catalyzes the reaction. Nextera® XT uses a mutated form of the Tn5 transposase, which has a higher activity than the wild type enzyme (Adey et al. 2010). This Nextera® XT enzyme fragments the DNA template and covalently tags the 5' ends with Illumina® specific sequencing primer sequences that are on the transposon ("tagmenting") during a five-minute reaction (Figure 8A, B). To add the necessary index sequences and flow cell adapters to the DNA fragments, a 12-cycle Polymerase Chain Reaction (PCR) is performed. Indices can be used to separate fragments for analysis following the completion of the run (Metzker 2010). After incorporation of these indices and adapters, the library preparation process is complete (Figure 8C) (Adey et al. 2010).

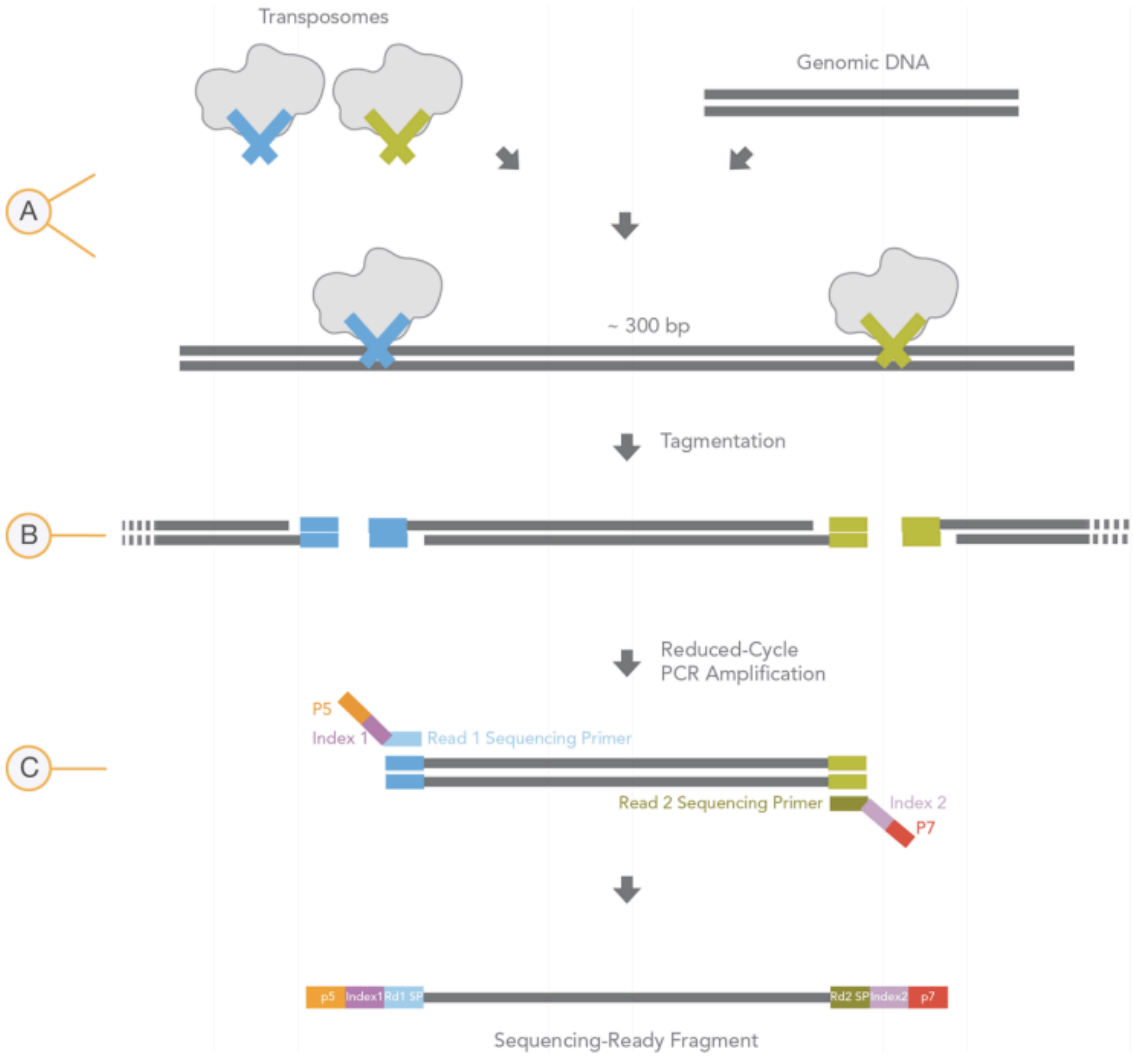


Figure 8: Illumina® Nextera® XT tagmentation and limited cycle PCR (Illumina® 2012). A: transposome with sequencing sites on the transposon ends (green and blue). B: DNA has been fragmented and tagged with sequencing sites. C: Top, limited-cycle PCR with primers specific to sequencing sites (blue and green), index (purple) and adapter sequences (orange and red). Bottom: completed DNA library.

After the 12-cycle PCR reaction, the DNA fragments are then size-selected with Agencourt® AMPure® XP beads. These paramagnetic beads select DNA molecules of different sizes based on altering the ratio of beads to DNA molecules; a higher ratio changes the electrostatic properties of the beads and thus allows for the binding of smaller DNA molecules, and vice versa (DeAngelis, Wang, and Hawkins 1995). Then,

the libraries are normalized in concentration with another magnetic bead-based technique, which attempts to ensure that all samples generate approximately the same amount of sequencing data. The DNA is bound to the beads, washed and eluted, after which the libraries can be pooled and sequenced (Illumina® 2012).

Advantages of the Illumina® Nextera® XT system are that it can be used to prepare complete DNA libraries from samples containing 1 nanogram of DNA in under four hours (Illumina® 2012). It follows a simple protocol with a significantly smaller amount of hands-on time than traditional library preparation methods.

With other methods, manual size selection with gel electrophoresis and quantification of the library prior to sequencing is necessary. Presently, the bead-based size selection in Nextera® XT does not require any additional visualization, although it is recommended that single-stranded DNA be quantified prior to sequencing.

Since indices are added to both sides of the DNA fragment, several samples may be multiplexed and run on the Illumina® MiSeq™ system simultaneously. As such, high-coverage data, which is necessary for reliable minor variant and mixture detection, can quickly be obtained. Thus, Illumina® Nextera® XT would make a useful system to employ in forensic laboratories, which need to sequence samples quickly, efficiently and accurately.

1.4 Bioinformatics

Bioinformatics, the science of collection and analysis of all biological data with computer software, is an important part of NGS since appropriate data handling is crucial

to proper interpretation. During sequencing, Illumina®'s Real Time Analysis (RTA) software performs base calling. In addition, Sequencing Analysis Viewer (SAV) outputs certain quality metrics of the sequencing run, such as the number of clusters generated, the intensity of the fluorescent signal, and quality (Q) scores of each base. These Q-scores are based on a logarithm of error probability at a certain position (Cock et al. 2010). For example, a score of Q30 indicates a 0.1% probability of an incorrect base call; lower Q scores indicate lower base accuracy (Table 1) (Minoche, Dohm, and Himmelbauer 2011).

Table 1: Quality scores as pertaining to base accuracy.

Quality score	Error probability	Base accuracy
10	1/10 bases	90%
20	1/100 bases	99%
30	1/1,000 bases	99.9%
40	1/10,000 bases	99.99%

After completion of the sequencing process, MiSeq™ Reporter (MSR), an on-board software package, performs secondary analysis (Illumina® 2013a). MSR demultiplexes the data, which means that the pool of sequence data from a run is separated by reading the short index sequences and then placed into distinct bins. FASTQ files are then generated, which contain both base calling information as well as the corresponding Q-scores for each base (Cock et al. 2010). In the case of the resequencing workflow in MSR, which is used for sequencing targets with a known reference sequence, the sequence reads in the FASTQ files are aligned to the reference.

In this workflow in MSR, alignment is performed by the Burrows-Wheeler Aligner (BWA) which aligns short reads to a reference while allowing short gaps and

mismatches (Li and Durbin 2009). BWA is a global aligner, which considers the complete read for successful alignment. Additional alignment algorithms are available in MSR including a banded Smith-Waterman algorithm. This is a local aligner, which considers different fragment sizes of each read during alignment.

If a reference sequence is used, then variants from the reference are called in the final step of the MSR workflow. In MSR, this is performed with either Genome Analysis Toolkit (GATK) (McKenna et al. 2010) or with Illumina®'s Somatic Variant Caller, which is able to detect low-level variants from a reference if proper thresholds are specified.

In addition to MSR, there are many other options available to analyze NGS data. Numerous bioinformatics tools are currently available, e.g. the Blat-Like Fast Accurate Search Tool (BFAST) for alignment (Homer, Merriman, and Nelson 2009). These tools can be installed directly onto a computer and controlled by typing commands into the command line interface (CLI). These command line tools give the user a very broad assortment of settings to design a custom pipeline suited to the type of information that needs to be gleaned from their sequence data (Kumar and Dudley 2007).

Integrative Genomics Viewer (IGV) is a Java-based freeware package developed by the Broad Institute, which provides the user with a Graphical User Interface (GUI) for simple implementation of commands. IGV can be used to view Sequence Alignment/Map (SAM) or Binary SAM (BAM) files, which are used to store read alignments in a smaller file size (Li et al. 2009), or Variant Call Format (VCF) files, which consist of variants from the rCRS extracted from an alignment file (Danecek et al. 2011). This software is particularly useful for scrolling through large sets of reads.

Visualization of data can be very helpful; IGV is fast, which allows the user, among other things, to observe the overall coverage throughout the genome, zoom in on the sequence to the nucleotide level and identify the location of low-level variants (Figure 9).

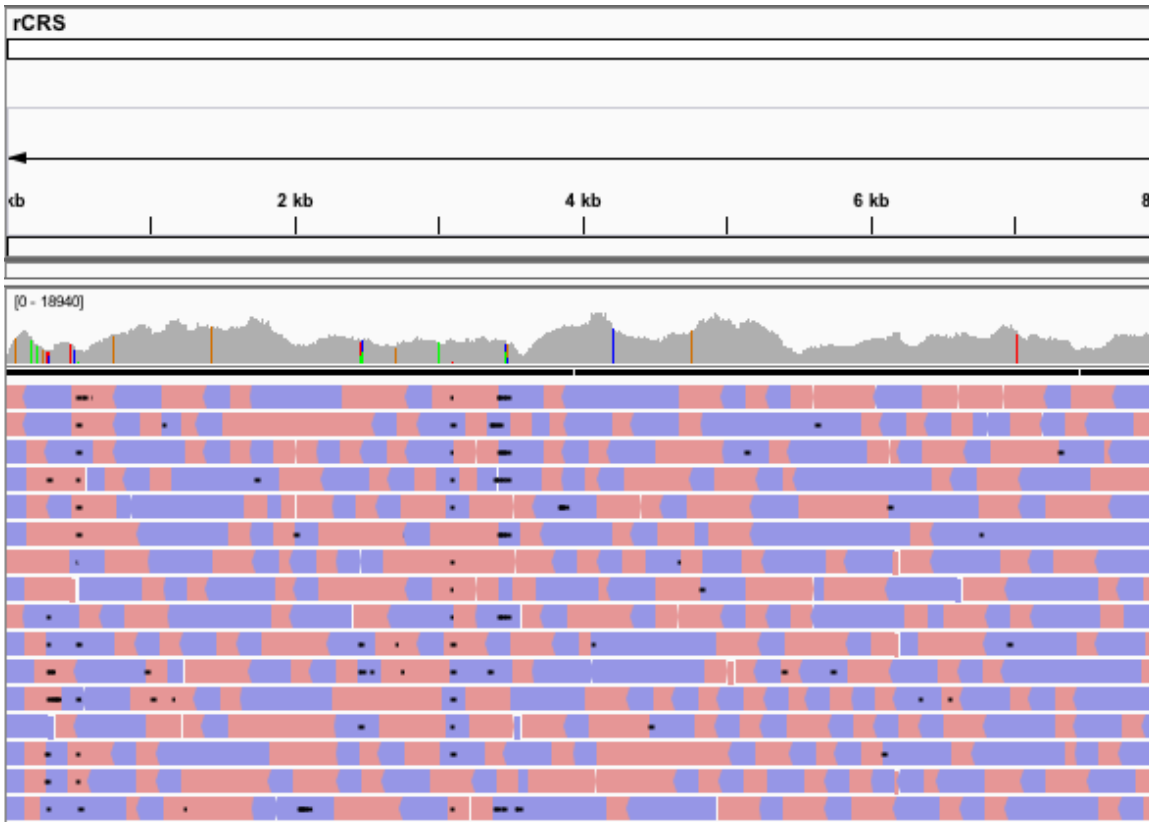


Figure 9: Visualization of a whole human mitochondrial genome dataset in IGV. Shown is sequence from base 1 to 8,000. Coverage is shown in gray, with variants from the reference sequence marked as narrow colored bars. Below, forward and reverse reads are indicated in red and blue, respectively, with deletions marked as black dashes.

Commercial software packages, such as CLC bio® CLC Genomics Workbench (CGW), offer a complete GUI-driven solution from importing a FASTQ file to creating alignments and calling variants. Compared to using command line tools, this commercial software package may limit control over parameters to simplify data analysis and increase user-friendliness. In addition, CGW is installed locally which, if a high-performance computer is used, can provide a tremendous increase in computing power

over cloud-based analysis software that depend on internet speed (CLC bio 2012). The integrated workflow option in CGW allows the user to create and save analysis pipelines so the same set of analyses can quickly be applied to several data sets. These features make CGW a powerful software package.

CGW encompasses many tools for different types of analyses, including Sanger sequencing and NGS of both RNA and DNA. The CGW software can be used to output extensive reports on data quality and metrics, and features a number of methods to visualize data. For NGS analysis CGW incorporates a custom alignment algorithm that supports local alignment. Variant calling can be performed with the Quality-based Variant Detection (QVD) tool, which calls variants from a reference based on the quality scores of the putative variant and its surrounding sequence. In contrast, the Probabilistic Variant Detection (PVD) tool calls variants based on a custom algorithm that combines Bayesian statistics and an estimation of Maximum Likelihood.

1.5 Whole Mitochondrial Genome Sequencing

As mentioned previously, Sanger sequencing is a relatively labor-intensive as well as expensive tool to analyze entire mtGenomes. NGS allows for sequencing of multiple human mtGenomes by massively parallel sequencing of short fragments of DNA (Pareek, Smoczynski, and Tretyn 2011). Thus, analysis of the mtGenome can be expanded outside of the HV regions, which encompass less than 4% of the entire molecule (Coble et al. 2004). Sequencing the whole mtGenome provides the forensic analyst with more data, which increases the exclusionary power of mtDNA. For example, from a group of 60 individuals that exhibited no or a single SNP variant from the rCRS in

the HV regions, 80% of individuals could be resolved when the analysis was expanded to the entire mtGenome (Andréasson et al. 2007). However, mtDNA population databases like EMPOP (Parson and Dür 2007) need to be broadened to incorporate whole genome reference sequences to obtain robust estimates of mtGenome rarity from casework analyses (Irwin et al. 2011).

This study is focused on the analysis of mtDNA in two distinct sample types:

1. Buccal swabs, which usually contain pristine DNA, can be easily amplified using a "targeted" PCR, or amplification performed on a specific region of the genome. The approach described in this study employs longer primer sets to encompass the entire mtGenome, which can be rapidly sequenced with NGS to generate data for reference purposes or the analysis of known samples.
2. Hair shaft extracts, a challenging sample type that often contains degraded DNA, are traditionally amplified with multiple shorter primer sets. For whole mtGenome sequencing, this can pose a limitation as these hair shaft extracts contain low concentrations of mtDNA. To reliably sequence hair shaft extracts with NGS, mtDNA in these samples may be pre-amplified, meaning that the total DNA template in a sample is amplified using a non-specific approach. The targeted PCR and pre-amplification techniques mentioned above will be discussed in detail in the following sections.

1.5.1 Long PCR - Targeted Amplification

The long PCR (LPCR) performed in this study is a targeted amplification that employs mtDNA specific primers to ensure mtDNA in the reference sample is amplified and available for NGS. Two primer sets were designed to create two amplicons of 9.1 and 11.1 kb in two separate reactions; these amplicons overlap at the control region in

order to generate double sequence coverage in this region (Figure 10). This LPCR requires largely intact DNA, and thus can be used on robust sample types such as blood or buccal swabs to generate whole mtGenome reference sequences. A similar dual-primer set long PCR approach has successfully been implemented on high quality mtDNA extracted from buccal swabs (Gunnarsdóttir et al. 2011) and blood samples (Fendt et al. 2009).

The TaKaRa™ Long and Accurate system was chosen to perform this LPCR amplification. TaKaRa™ consists of a mixture of a *Taq* polymerase and a proofreading polymerase with 3′-5′ exonuclease activity. This system has a fidelity 6.5 times higher than conventional *Taq* polymerase, and is routinely used to generate amplicons up to 25, (Goto, Nishino, and Hayashi 2006) and amplicons as large as 50 kb have been reported (Seki, Hayashida, and Shinozaki 1996).

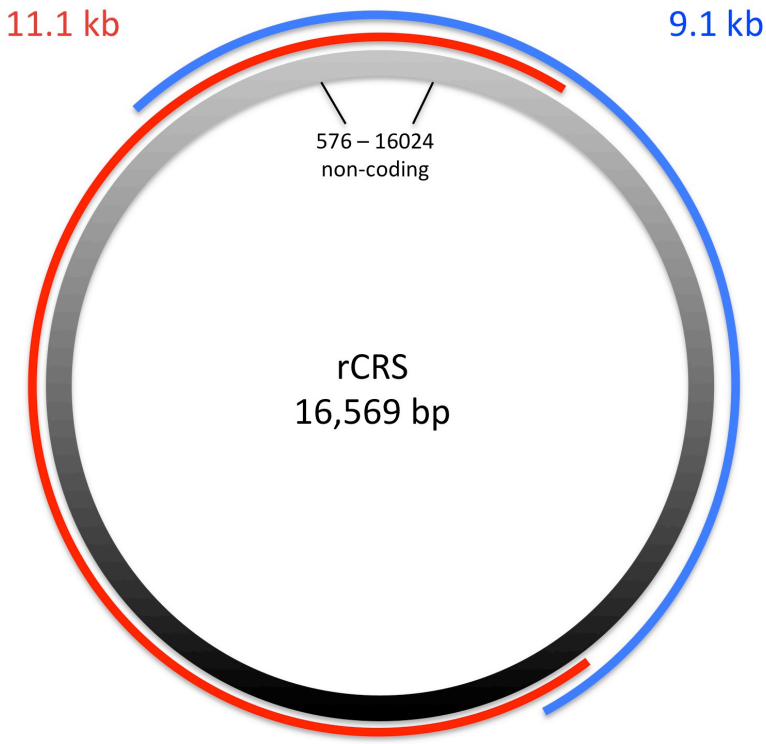


Figure 10. Long PCR performed with two primer sets. This results in a 9065 bp amplicon and an 11170 bp amplicon.

Conventional *Taq* polymerases show a higher rate of base misincorporation than other types of polymerases. If base mismatches occur at priming sites, these may cause the polymerase to stall during extension (Huang, Arnheim, and Goodman 1992, Barnes 1994). Because the proofreading enzyme in the TaKaRa LA Taq™ system removes 3' bases, this *Taq* polymerase is less likely to stall, and thus longer amplicons can be generated (Davies and Gray 2002).

1.5.2 Whole Genome Amplification - Pre-amplification

Hairs are a typically challenging sample type found in forensic casework. The amount of cell nuclei in a hair shaft is extremely limited and can vary from individual to individual, which adds to the difficulty of extracting nuclear DNA from hairs (Szabo et

al. 2012). Thus, it is more feasible to extract small quantities of mitochondrial DNA from hair shafts. These extracts often contain degraded mtDNA, and they therefore require amplification with shorter primer sets to successfully generate an amplicon. As hair shaft extracts are often limited in mtDNA copy number, it is less feasible to generate whole mtGenome sequences from these samples by PCR amplification alone. Potentially, Whole Genome Amplification (WGA), a technique initially developed for random amplification of the entire genome (Dean et al. 2002) can overcome this issue by creating more DNA template from an extract. This pre-amplified material may then be used to generate more PCR amplicons around the mtGenome. These amplicons can then be sequenced to generate whole mtGenome NGS data from challenging sample types.

Some WGA methods are performed at a constant temperature (isothermal), whereas others are based on PCR. Multiple Displacement Amplification (MDA) is one of several isothermal WGA methods. This technique employs a Ø29 bacteriophage DNA polymerase to extend random hexamer oligos during an isothermal reaction. This polymerase has 3' -> 5' proofreading activity and is capable of generating DNA strands of 70,000 base pairs and higher (Blanco et al. 1989). Some MDA methods can yield up to a 100,000-fold increase in genomic DNA.

During an MDA reaction, random hexamers anneal to the template DNA and are extended (see Figure 11). Upon encountering a previously extended strand, the Ø29 polymerase displaces that strand and continues copying the template DNA. Ergo, the 3' ends of the strands displace 5' ends of other, adjacent, strands that are extended in the same direction. Primers anneal to these newly synthesized strands and are extended in the reverse direction as well. This process results in the creation of a branched network of

amplified DNA molecules (Schneider et al. 2004). As stated previously, as this is not a targeted amplification, the generated DNA template is non-specific.

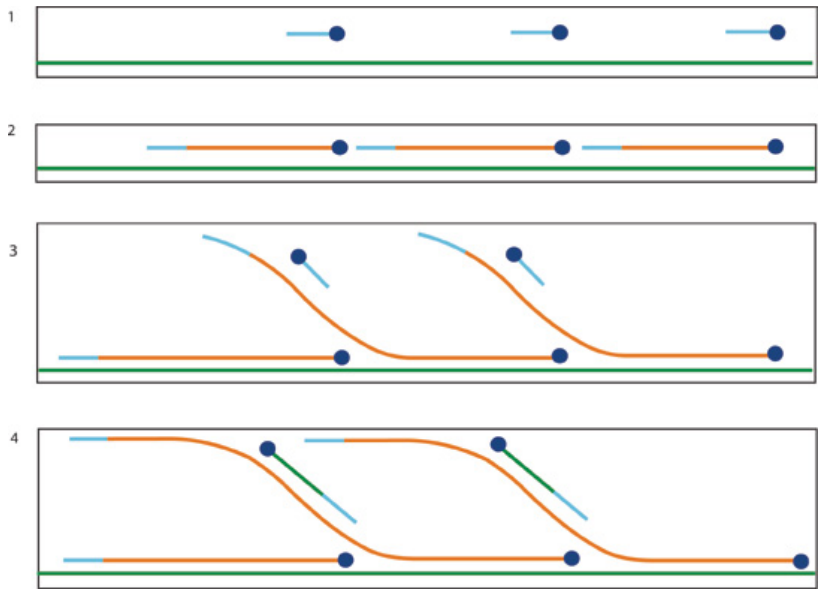


Figure 11: Strand displacement during whole genome amplification (Spits et al. 2006). 1-2: Primers anneal randomly and are extended. 3-4: Extending strands displace existing strands, while additional hexamer primers anneal and are extended in parallel.

Sigma-Aldrich® GenomePlex® is a PCR-based WGA method. This method incorporates a series of steps to fragment the DNA template, after which universal adapters are ligated to the fragments. These adapters subsequently facilitate PCR amplification with a single universal primer (Figure 12). Traditional MDA reactions have shown to exhibit an amplification bias (preferential amplification of certain loci over others) less than 3-fold. This is significantly lower than many PCR-based WGA methods, which have exhibited amplification bias up to $10^3 – 10^6$ fold (Dean et al. 2002). However, the GenomePlex® method, as opposed to other PCR-based WGA methods, has not shown to introduce any significant amplification bias when compared to an MDA method (Barker et al. 2004).

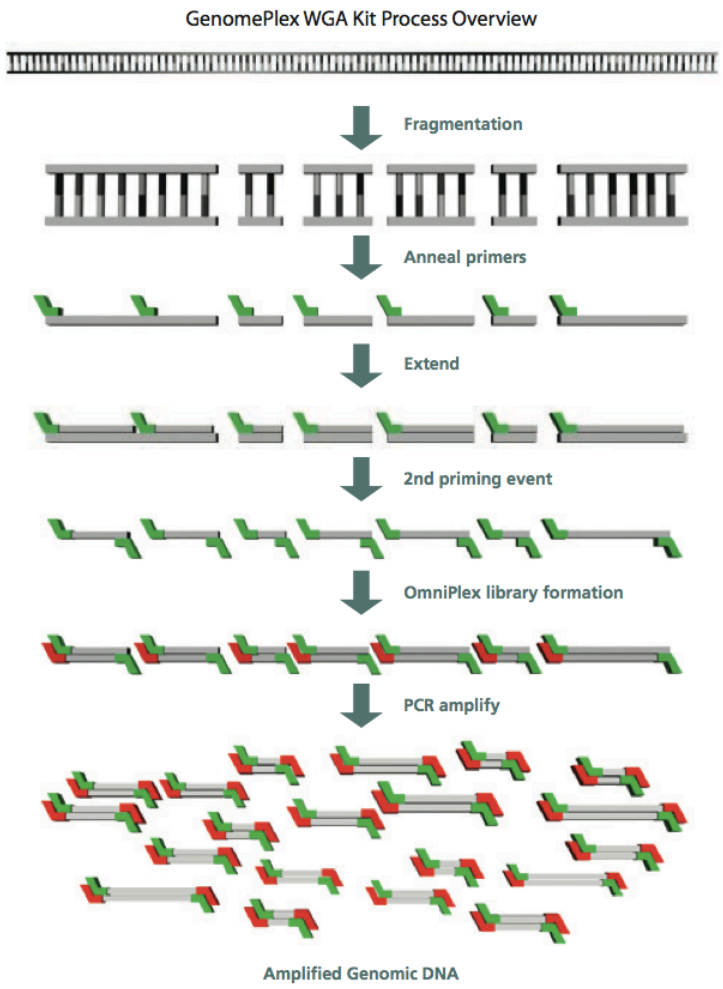


Figure 12: Overview of the GenomePlex® technique (Sigma-Aldrich®). DNA is fragmented, after which primers are annealed and extended. The product is then amplified with a single primer.

WGA techniques should be evaluated in two categories: efficiency and accuracy.

1. Efficiency: in order to positively affect downstream processes the method should sufficiently amplify the starting material in hair shaft extracts. It has been shown that some WGA methods do not perform well in this respect when limited starting material is present. In some cases, the primers have been known to branch off each other (hyperbranching), which creates large quantities of non-specific,

primer-derived material (Barber and Foran 2006, Lage et al. 2003). This can cause issues in downstream analyses, both in sequencing (low amounts of target sequence present) as well as quantification (overestimation of target material concentration when using non-specific quantification methods).

2. Accuracy: the method used should accurately amplify the starting material. Due to the high clonal amplification by WGA methods, base misincorporations are a concern, especially with the introduction of very sensitive NGS technologies. Even small base misincorporations may ultimately affect the mtDNA interpretation in casework.

In this research effort, four different commercially available WGA kits were evaluated to compare their efficacy and accuracy in amplifying mtDNA from hair shaft extracts. Due to its high amplification efficiency and accuracy, the MDA technique is promising, as shown in studies cited above. It has been noted that the MDA technique does not seem to perform as well on degraded DNA, due to a lower probability of primer annealing, as binding sites may have been fragmented (Barber and Foran 2006). In addition, primers are not likely to consistently bind near the ends of the fragments. Thus, primer extension may e.g. start in the middle of a DNA fragment, which creates an extension product half the size of the original fragment. As the DNA may be further fragmented during amplification, the activity of the enzyme is not optimally utilized. However, this study employs an optimized hair extraction protocol, which may improve the quality of extracted DNA and potentially facilitate more efficient amplification using the MDA method. Moreover, even a small fold amplification may be sufficient for effective sequencing using NGS, since the Illumina® Nextera® XT library preparation

technique requires only 1 ng input of double-stranded template DNA.

Three isothermal MDA kits were used in this study: the QIAGEN® REPLI-g Mini kit, Mitochondrial DNA kit and Single Cell kit. The Mini kit is optimized for >1 ng of input DNA, whereas the Mitochondrial DNA kit employs mtDNA-specific hexamers in addition to random hexamers, and the Single Cell kit is optimized for sample inputs as low as a single cell (QIAGEN® 2011a, QIAGEN® 2011b, QIAGEN® 2012a).

One PCR-based kit was chosen to compare the efficiency of a PCR-based method to MDA methods. The Sigma-Aldrich® GenomePlex® kit was selected since it is one of a few PCR-based WGA methods that has not shown to introduce any significant amplification bias when compared to an MDA method (Barker et al. 2004). Furthermore, in a comparison between seven different WGA kits, the GenomePlex® kit was deemed to be the best when using degraded starting material, since mtDNA typing subsequent to WGA was successful even for DNA that was fragmented to 100 bp (Maciejewska, Jakubowska, and Pawłowski 2013). Additionally, this kit may be useful since the initial fragmentation step in the protocol can be omitted in cases of potential fragmentation of the DNA template, such as with hair shaft extracts. This may theoretically generate larger DNA fragments than the MDA method applied to degraded samples, as explained previously.

1.6 Mitochondrial DNA Quantification

Assessing the quantity of mitochondrial DNA in samples is critical to this study; by quantifying the amount of mtDNA in hair samples before and after WGA, it is

possible to calculate how well WGA has pre-amplified the mtDNA in these samples. This was assessed using real-time PCR, with a sensitive 5' nuclease assay specific to human mtDNA (Kavlick et al. 2011). Real-time PCR (rtPCR or qPCR) is a technique often used to quantify DNA. The qPCR performed in this study is a 5' nuclease assay, which employs primers and probes designed for a specific target sequence (Heid et al. 1996).

Probes are short target-specific DNA sequences which hybridize to the template DNA between the forward and reverse primer during the annealing step of the qPCR (Heid et al. 1996). A nonfluorescent quencher (NFQ) and a minor groove binder (MGB) are attached to the 3' end of the probe, and a fluorescent reporter dye is attached at the 5' end. When the probe is bound to the template, the reporter and quencher are in close proximity, which prevents the reporter dye from emitting fluorescence. However, during primer extension the polymerase degrades the probe, which releases the reporter dye and results in fluorescence.

Since this is a PCR-based technique, the DNA template accumulates during amplification. A CCD camera in the qPCR instrument is used to detect the increasing amount of fluorescence generated by the quencher. When this fluorescence crosses a predetermined threshold, the instrument notes the cycle number during which this occurs and assigns a cycle threshold (C_T) value to the sample. The C_T values for the samples can be compared to those of the standard curve, which allows for determination of the mtDNA copy number in a sample. As with all 5' nuclease assays, the human mtDNA qPCR assay used in this study is sensitive, and can accurately detect a quantity of mitochondrial DNA in a sample ranging from 10 to 10^8 copies (Kavlick et al. 2011). QPCR also applies an internal positive control (IPC) assay that uses separate IPC

primers, DNA and a probe to verify whether the reagents and the instrument are working properly and to detect the presence of inhibitors (Honeycutt, Sobral, and McClelland 1997).

1.7 Objectives

The present research effort studies the preparation of mtDNA in reference and forensic sample types for NGS, and aims to fulfill three goals:

1. To establish a method for the targeted amplification and NGS of mtDNA from reference samples. Buccal swabs usually contain pristine DNA, which can be amplified using a long PCR approach and rapidly sequenced with NGS to establish whole mtGenome reference sequence for comparison purposes. Due to the amount of data that can be obtained with NGS, many samples can be run at once, depending on how many indices are employed.

2. To evaluate the use of different Whole Genome Amplification techniques on forensic sample types such as hair or bone. These are traditionally more difficult to analyze, since they often contain degraded DNA and are limited in mtDNA copy number. Therefore, these samples may require amplification with shorter primer sets. Potentially, WGA can overcome this limitation by creating more DNA template molecules from the same sample. In addition, with NGS the limit of detection of minor variants in mixtures is lowered compared to that of Sanger sequencing, allowing an enhanced detection of mixed samples.

3. To evaluate Illumina® Nextera® XT technology for mtDNA library preparation and subsequent Illumina® MiSeq™ sequencing in a forensic context. Library

preparation is an important step in NGS, since the appropriate amount of clusters need to be generated for a run to be successful. In addition, library preparation should allow for accurate sequencing of DNA samples. For use in forensic laboratories, the method should be simple and fast and should allow for the multiplexing of several whole mtGenomes in one run while generating high-coverage data. The Illumina® Nextera® XR method was evaluated by preparing different sample types for sequencing; WGA product, of which a subset was additionally PCR amplified prior to sequencing, and long PCR products were prepared with this kit and sequenced on the Illumina® MiSeq™ in two separate sequencing runs. This study fills a gap in present knowledge, as prior to this research no studies have been published that combines this particular NGS library preparation technique with forensic human mtDNA analysis.

CHAPTER 2: MATERIALS AND METHODS

The flow chart below (Figure 13) illustrates the two different pathways that were taken in this study for analysis of reference samples and challenging samples.

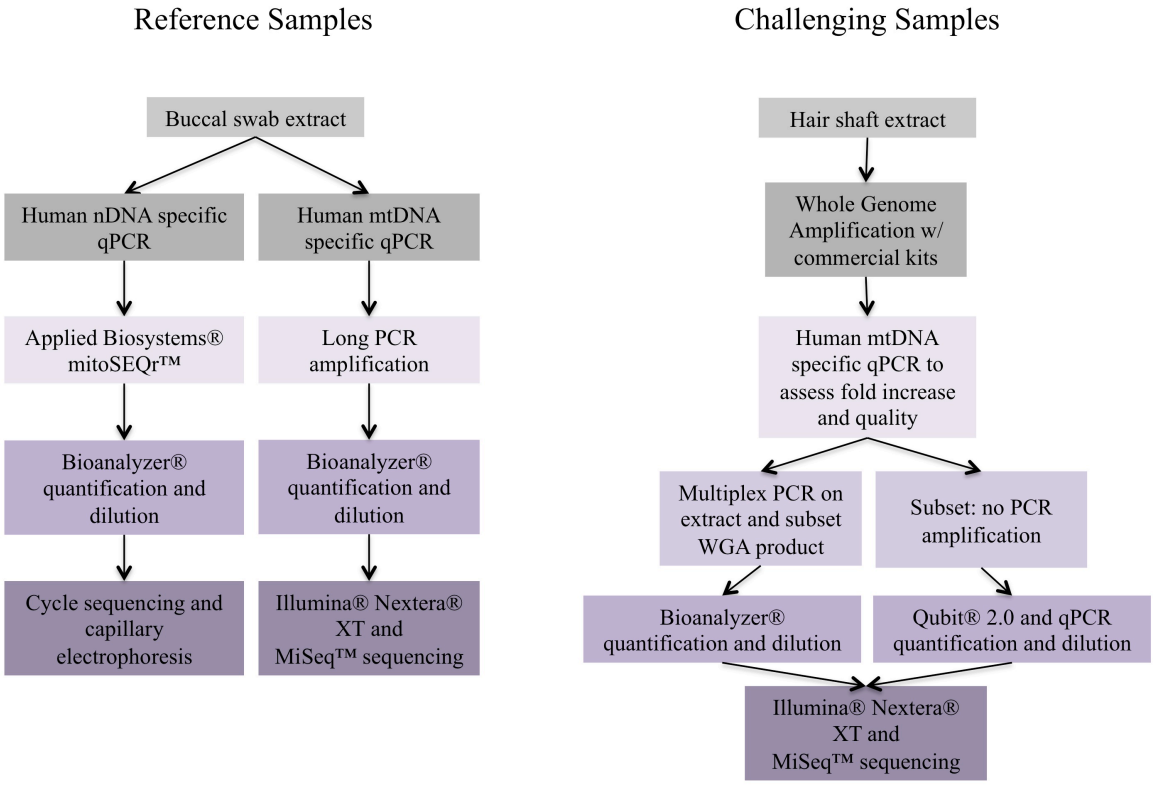


Figure 13: Experiments performed for reference samples and challenging samples.

MtDNA research is highly prone to contamination, and therefore requires the analyst to maintain pristine laboratory conditions. For each experiment, all objects in the laminar flow hood were cleaned with 10% bleach and 70% ethanol, after which the hood was UV irradiated for 15 minutes. Any reagents that do not hold biological material, such as sterile water and TE buffer, as well as any other items necessary for the experiment

such as 1.5 ml Eppendorf tubes, optical plates or caps, were UV irradiated for 15 minutes either in the laminar flow hood or in a Spectrolinker® XL-1000 UV Crosslinker. During the experiment, the analyst's hair was tied back and they wore a laboratory coat, facemask, a pair of gloves under a set of disposable sleeves and a pair of gloves over it. The outer gloves were renewed upon suspected contamination.

2.1 Collection of Reference Material

Buccal swabs, hairs and whole blood delivered directly on Whatman™ FTA™ cards were obtained from eight donors according to approved Institutional Review Board (IRB) protocol.

DNA was extracted from the buccal swabs using the Qiagen® QIAamp® DNA Mini Kit™ (QIAGEN® 2012b) according to the manufacturer's instructions, and eluted in 150 µl TE buffer.

From the FTA™ cards, 1.2 mm punches were taken with a Whatman™ Harris Micro Punch, which were then washed using vendor's protocol (Whatman), for a total of eight washes with Whatman™ FTA™ Purification Reagent.

2.2 Creation of Reference Sequences with Sanger Sequencing

DNA extracted from buccal swabs, as described in section 2.1, was used to generate whole mtGenome reference data using the Applied Biosystems® mitoSEQr™ assay, which employs 46 primer pairs to PCR amplify the entire mtGenome (Applied Biosystems® 2006). The concentration of nuclear DNA in these extracts was quantified

with the Applied Biosystems® Quantifiler® Human DNA Quantification kit on an Applied Biosystems 7500 Real-Time PCR System according to vendor's protocol (Applied Biosystems®). The extracts were then diluted to 1 ng/µl.

2.2.1 Amplification with AmpliTaq Gold®

To PCR amplify the DNA in the donor samples, reactions were set up as shown in Table 2, for each of the 46 primer pairs.

Table 2: Reaction conditions for the Applied Biosystems® mitoSEQr™ assay with AmpliTaq Gold®

Reagent	**µl/reaction**	**Final concentration**
DNA Template (1 ng/µl)*	1	-
Primer Pair	2	60 nM each
AmpliTaq Gold® PCR Master Mix (2x)	5	1x
50% glycerol	1.6	5%
Sterile water	0.4	-
Total volume	**10**	**-**

*In addition to 1 ng/µl diluted buccal swab extracts, for the amplification of five amplicons across two donors 1.2 mm punches from whole blood samples stored on FTA™ cards were used. In these cases, 1 µl of additional sterile water was added to the reaction mixture, as FTA™ punches are not considered to add volume to the PCR reaction.

The same reaction conditions were used for the extraction reagent blank and a positive control. For the reagent blank, water was taken through the extraction process, while for the positive control, 1 ng of purified genomic DNA from Applied Biosystems® was amplified with one randomly chosen primer pair from the set of 46 and one primer pair that was held consistent throughout all amplifications. This genomic DNA was provided to the Forensic Science program at Western Carolina University by Applied Biosystems® and is not commercially available. Amplification was performed under the thermal cycling conditions shown in Table 3 using an Applied Biosystems® Veriti® 96-Well thermal cycler.

Table 3: Thermal cycling conditions for the Applied Biosystems® mitoSEQr™ assay with AmpliTaq Gold®

96°C for 5 min	
94°C for 30 s	40 cycles
60°C for 45 s	
72°C for 45 s	
72°C for 10 min	
4°C hold	

2.2.2 Amplification with Roche FastStart™

Repeated amplification issues were seen with approximately 36 primer sets across five donors. Troubleshooting these issues by changing the DNA template from buccal swabs to punches from blood FTA cards, as well as extracting fresh buccal swabs, allowed for successful amplification of some, but not all, amplicons. For the remaining 12 primer sets across four donors it was decided to alter the master mix. For these 12 primer sets the AmpliTaq Gold® master mix was substituted with Roche FastStart™ reagents (Table 4).

Table 4: Reaction conditions for the Applied Biosystems® mitoSEQr™ assay with Roche FastStart™. Conditions were altered as compared to Table 2, due to master mix requirements for the FastStart™ enzyme.

Reagent	**µl/reaction**	**Final concentration**
FastStart High Fidelity Reaction Buffer 10x, with 18 mM $MgCl_2$	2.5	1x, 1.8 mM $MgCl_2$
DMSO	2.5	10 %
dNTP Mix, 10 mM each	0.5	200 µM each
FastStart High Fidelity Enzyme Blend	0.25	1.25 U
Primer mix	8.4	0.2 µM each
DNA template (1 ng/ul)	1	-
Sterile water	9.85	-
Total volume	**25**	**-**

Again, the same conditions were used for the extraction reagent blank and 1 ng of Applied Biosystems® genomic DNA. Amplification was performed as per Table 5 on an Applied Biosystems® Veriti® 96-Well thermal cycler. Due to the expected higher activity of the Roche FastStart™ enzyme as compared to the AmpliTaq Gold® master mix, the number of PCR cycles was lowered in comparison to the usual mitoSEQr™ protocol.

Table 5: Thermal cycling conditions for the Applied Biosystems® mitoSEQr™ assay with Roche FastStart™

95°C	2 min	
95°C	30 sec	35 cycles
60°C	30 sec	
72°C	45 sec	
72°C	7 min	
4°C	hold	

2.2.3 PCR Purification and Cycle Sequencing

Quantification of the amplicons was performed with the Agilent Technologies® 2100 Bioanalyzer® using the Agilent Technologies® DNA 1000 Kit™ (Agilent Technologies 2006a). Primers and unincorporated nucleotides were degraded by adding 2 µl of USB® ExoSAP-IT® to 5 µl of the PCR reactions; this mixture was incubated 15 min at 37°C followed by 15 min at 80°C (Affymetrix® 2011). The concentration of PCR product after USB® ExoSAP-IT® was recalculated, to account for the dilution introduced by the enzyme treatment.

Cycle sequencing was performed with Applied Biosystems® BigDye® Terminator v1.1 Ready Reaction Mix (Applied Biosystems® 2010), which was diluted in a 1:4 ratio with sterile water. In cases where the recalculated concentration of PCR product was

higher than 2.86 ng/μl, a half-volume reaction with diluted BigDye® reaction was performed, otherwise a full-volume reaction was performed (Table 6).

Table 6: Cycle sequencing reaction conditions with diluted Applied Biosystems® BigDye® Terminator v1.1 Ready Reaction Mix

Reagent	**Half-volume Reaction (>2.86 ng/μl)**	**Full-volume Reaction (<2.86 ng/μl)**
BigDye, diluted 1:4	4.75 μl	9.5 μl
M13 Primer (Forward or Reverse), 0.56 μM final concentration	1.75 μl	3.5 μl
Template	10 ng	Up to 10 ng
Sterile water	3.5 μl	7 μl

For the pGEM sequencing control, which is provided with the BigDye® kit, 3.5 μl of DNA (0.2 μg/μl) was used in half-volume reactions and 5 μl was used in full-volume reactions. Thermal cycling was performed as shown in Table 7 on an Applied Biosystems® Veriti® 96-Well thermal cycler.

Table 7: Thermal cycling parameters for cycle sequencing with diluted Applied Biosystems® BigDye® Terminator v1.1 Ready Reaction Mix

96°C for 1 min	
96°C for 10 s	25 cycles
50°C for 5 s	
60°C for 4 min	
4°C hold	

2.2.4 Cycle Sequencing Purification and Capillary Electrophoresis

Purification of the cycle sequencing product was performed with Agencourt® CleanSeq® beads (Agencourt®). For half-volume reactions, 5 μl of reaction was purified using 10 μl of beads and 31 μl of 85% ethanol. For full-volume reactions, 10 μl of

reaction was purified using 10 µl of beads and 42 µl of 85% ethanol. The DNA was eluted in 40 µl of 0.1 mM EDTA and 30 µl was sequenced on the Applied Biosystems® 3130*xl* Genetic Analyzer with POP-6™ polymer. Base calling was performed with Applied Biosystems® Sequencing Analysis 5.2 and Gene Codes Sequencher® 5.0 was used to identify variants from the rCRS.

2.3 Mitochondrial DNA Quantification

The quantity of mtDNA in samples was assessed with the human mitochondrial DNA specific qPCR assay mentioned in chapter 1.4, as described by Kavlick *et al* (2011).

After distribution of the master mix and controls, wells of the optical plate were loosely capped with optical caps and covered with cross-linked aluminum foil to minimize photobleaching of the reagents. Then, the standard dilution series was created and pipetted in duplicate into its respective wells. Outer gloves were changed after each standard. When possible, samples were pipetted in triplicate and wells were capped, after which the plate was centrifuged to consolidate the reagents. Plates were run on an ABI PRISM® 7000 Sequence Detection System with Sequence Detection Software v. 1.2.3 or an Applied Biosystems® 7500 Real-Time PCR System with HID Real-Time PCR Analysis Software v. 2.0.1.

2.4 Long PCR Amplification of Buccal Swab Extracts

To amplify the entire mtGenome in two reactions, two novel primer sets were designed to overlap at the HV region, to potentially double sequence coverage in these regions (Table 8).

Table 8: Long PCR primer sets. Set 1: 9,065 bp amplicon, Set 2: 11,170 bp amplicon.

1F 5' AAA GCA CAT ACC AAG GCC AC 3'
1R 5' TTG GCT CTC CTT GCA AAG TT 3'

2F 5' TAT CCG CCA TCC CAT ACA TT 3'
2R 5' AAT GTT GAG CCG TAG ATG CC 3'

These two primer sets were used to amplify approximately 200,000 copies of mtDNA as template in two separate reactions (Table 9) with the TaKaRa® LA Taq system described in section 1.3.1 (TaKaRa Bio Inc. 2013).

Table 9: Reaction conditions for Long PCR with TaKaRa® LA Taq®

Reagent	**µl/reaction**	**Final concentration**
DNA Template (200,000 copies mtDNA)	1	-
Forward primer	1	0.2 µM
Reverse primer	1	0.2 µM
10x TaKaRa® LA PCR buffer	5	1x
TaKaRa® LA dNTP mix (2.5 mM each)	8	0.4 mM each
TaKaRa® LA Taq® polymerase (5U/µl)	0.5	2.5U
Sterile water	33.5	--
Total volume	**50**	**-**

In addition, 1 ng of HL60 DNA was amplified as a positive control and 10 µl of sterile water was used as a negative control for both primer sets. Thermal cycling was performed as described in Table 10 on an Applied Biosystems® Veriti® 96-Well thermal cycler. Thermal cycling conditions were adapted from the vendor's protocol to incorporate a separate annealing step and to extend the extension time to 11 minutes, as the vendor recommends an extension time of 1 minute per kb of target DNA.

Table 10: Thermal cycling conditions for Long PCR with TaKaRa® LA Taq®

94°C	1 min	
94°C	30 sec	30 cycles
54°C	15 sec	
68°C	11 min	
72°C	10 min	
4°C	hold	

After amplification, the long PCR products were quantified using the Agilent Technologies® 2100 Bioanalyzer® using the Agilent Technologies® DNA 12000 Kit™ which is able to quantify DNA fragments of 100 - 12,000 bp in size (Agilent Technologies 2006b). Samples were purified with the Zymo® Clean & Concentrator-5™ kit using a 2:1 v/v ratio of DNA binding buffer to PCR product (Zymo Research) and requantified with the Agilent Technologies® DNA 12000 Kit™.

2.5 National Institute of Standards and Technology Standards for Sequencing

The National Institute of Standards and Technology (NIST) provided two sets of sequencing standards: Standard Reference Material (SRM) 2392 and 2394. SRM 2392,

Mitochondrial DNA Sequencing Standard (Human), is comprised of DNA extracts of the highly characterized lymphoblastoid cell lines 9947A and CHR as well as the cloned HV1 region from CHR (Levin, Cheng, and Reeder 1999). SRM 2394, Heteroplasmic Mitochondrial DNA Mutation Detection Standard, is comprised of a 285 bp amplicon amplified from CHR and 9947A, which differ by a single base pair at the same nucleotide position. These two amplicons have been mixed at ten defined ratios ranging from 1 to a 100% (Hancock, Tully, and Levin 2005). The SRMs were prepared for NGS alongside the LPCR amplicons as detailed in section 2.6. They were sequenced without any PCR amplification, to assess their value for use as sequencing controls.

2.6 Illumina® Nextera® XT and Sequencing on Illumina® MiSeq ™

In total, 25 samples were processed with Illumina® Nextera® XT: long PCR products of eight donors, three reagent blanks for the corresponding DNA extractions and an HL60 positive control, as well as all three SRM 2392 standards and all ten SRM 2394 standards which were not PCR amplified prior to sequencing.

The NIST standards were diluted to 200 pg/μl with sterile water, and 5 μl (1 ng) of each standard was pipetted into a separate well of a 96-well plate. For the LPCR samples, the 11.1 kb amplicons were diluted to 200 pg/μl and the 9.1 kb amplicons were diluted to 162 pg/μl with sterile water, after which 2.5 μl of both were pooled into a well of the 96-well plate, for a total of 0.905 ng which is slightly less than the recommended input. Due to the lower molecular weight of the 9.1 kb amplicon compared to the 11.1 kb amplicon, the 9.1 kb amplicons were diluted to a lower concentration than the recommended 200

pg/µl. This was performed in an attempt to equalize the number of 9.1 kb and 11.1 kb fragments for each donor, which may balance sequence coverage throughout the mtGenome. For each reagent blank, 2.5 µl of each PCR reaction (9.1 and 11.1 kb) was pooled undiluted into a separate well of the 96-well plate.

Tagmentation was performed on an Applied Biosystems® Veriti® 96-Well thermal cycler and followed by neutralization, after which each sample was assigned a unique index combination according to Illumina® guidelines (Illumina® 2012). Indexes and adapters were incorporated during a limited-cycle PCR amplification, which was performed on the Veriti® thermal cycler. Libraries were stored on the thermal cycler at 10°C overnight.

Purification of the PCR product was performed with a 0.6x ratio of Agencourt® AMPure® XP beads to PCR product, which is recommended by Illumina® when starting with DNA larger than 500 bp. Purification was immediately followed by library normalization with Nextera® XT magnetic beads. The libraries were then quantified with the Qubit® ssDNA Assay kit. All normalized samples were pooled into the Pooled Amplicon Library (PAL).

Illumina® PhiX Control, derived from the highly characterized phiX174 (RF1) bacteriophage which has a 5386 bp circular genome (Thermo Scientific), is commonly used as an Illumina® sequencing control (Kircher, Stenzel, and Kelso 2009). A 12.5 pM concentration of Illumina® PhiX v3 Control was spiked into the PAL at a 20% v/v ratio (Illumina® 2013b) prior to the 25-fold dilution of the PAL to create the Diluted Amplicon Library (DAL). A sample sheet was set up in Illumina®'s Experiment Manager to sequence all 25 samples in a Resequencing workflow with Nextera® XT as

the library preparation method and the rCRS as reference genome. Each sample and its corresponding indices were listed in the sample sheet, and the Somatic Variant Caller was specified for variant calling with a frequency cutoff of 0.001. The sample sheet was then uploaded into the MiSeq™ Control Software, which guides the user through loading of the reagents and allows for visualization of quality statistics during the progress of the run.

The DAL was sequenced on the Illumina® MiSeq™ in a 2x150 bp paired end v2 run. Sequencing analysis was performed with Illumina® Sequence Analysis Viewer 1.8, Illumina® MiSeq™ Reporter 2.2 and Integrative Genomics Viewer 2.2 and 2.3. NGS sequences were compared to those derived for the same donors with Sanger sequencing (section 2.1). Positions that did not exhibit a common base in this comparison of treatments were designated as sequence differences.

2.7 Whole Genome Amplification of Buccal Extracts

Initially, Whole Genome Amplification was performed on dilutions of buccal swab extracts, to determine if, and how well, WGA would amplify mtDNA in robust low copy number DNA samples.

DNA was extracted from buccal swabs as described previously in section 2.1. In a first experiment, DNA from one buccal swab extract was quantified and diluted serially to 100 pg/µl, 50 pg/µl, 25 pg/µl, 12.5 pg/µl, 6.25 pg/µl, 3.13 pg/µl, 1.56 pg/µl, 785 fg/µl, 392.5 fg/µl and 196.25 fg/µl of nuclear DNA. In the second, extracts from two donors were serially diluted to 100 pg/µl, 3.13 pg/µl and 196.25 fg/µl nDNA. For the third

experiment, the same two extracts were serially diluted, but to a specific mtDNA copy number: 20250, 6750, 2250, 750 and 250 copies/μl.

In each experiment the dilutions, their reagent blank, a negative control and a positive control were processed with four different kits according to their manufacturer's protocols, without any deviations. The QIAGEN® REPLI-g® Mini (QIAGEN® 2011a), Mitochondrial DNA (QIAGEN® 2011b) and the Sigma-Aldrich® GenomePlex® WGA2 kit (Sigma-Aldrich® 2012) were used with 5 μl of DNA input. The QIAGEN® Single Cell (QIAGEN® 2012a) kit requires 2.5 μl input.

In the first experiment, 10 ng of QIAGEN® REPLI-g® Human Control DNA (10 ng/μl) was used as a positive control. To preserve this positive control, in the second experiment 2.5 ng of the same control (diluted to 500 pg/μl) was used for the QIAGEN® REPLI-g® Mini kit and the Sigma-Aldrich® GenomePlex® WGA2 kit, but due to an oversight 1.25 ng was used for the REPLI-g® Mitochondrial DNA and Single Cell kits: the maximum input for the Single Cell kit (2.5 μl) was not taken into account. In the third experiment, 2.5 ng of the QIAGEN® control was used alongside 250 pg of an HL60 control, the latter being a more forensically relevant concentration.

Amplification took place on an Applied Biosystems® Veriti® 96-Well thermal cycler as well as several Applied Biosystems® GeneAmp® PCR System 9700 thermal cyclers. A qPCR quantification of mtDNA in the WGA product was performed as described in section 2.3. Quantification of pre-WGA and post-WGA product for each kit was performed on the same 96-well optical plate to minimize variation between quantifications. With these data, the fold-increases of mtDNA after WGA were assessed.

2.8 Whole Genome Amplification of Hair Extracts

After experiments with buccal extracts, Whole Genome Amplification was performed on DNA extracted from hair shaft.

2.8.1 Hair Shaft Extraction

One hair was obtained from each of three selected donors. DNA from these hairs was extracted following a newly developed hair extraction protocol (Burnside *et al*, 2012). The hairs were observed under a Fisher Scientific™ Stereomaster™ microscope at 25x magnification and the roots were removed. Two cm of hair nearest to the root was cut off and sonicated in 5% Alconox® Tergazyme™ for 20 minutes, after which the hair fragment was rinsed in 100% ethanol followed by a water rinse. The fragments were digested with QIAGEN® Buffer ATL, which was supplemented with proteinase K and dithiothreitol (DTT), at 56°C for one hour or until hairs were visibly digested. This digestion was followed by a brief incubation at 70°C with QIAGEN® Buffer AL (QIAGEN® 2010). The DNA was then purified using the Applied Biosystems® Prepfiler® Forensic DNA Extraction Kit (Applied Biosystems® 2012) and eluted in 50 - 60 µl of Prepfiler® Elution Buffer.

2.8.2 Whole Genome Amplification of Hair Shaft Extracts

Two duplicates of each hair extract, the reagent blank for the extraction process, a negative control and 2.5 µl of a 100 pg/µl HL60 positive control were amplified with the

QIAGEN® REPLI-g® Mini, Mitochondrial DNA and Single Cell kits as well as the Sigma-Aldrich® GenomePlex® WGA2 kit as described in section 2.7. However, since DNA from hair shaft extracts may already be fragmented, in two out of four cases the vendor's protocol for the Sigma-Aldrich® GenomePlex® WGA2 kit was modified to omit the fragmentation step, in an attempt to improve DNA quality for downstream applications. According to vendor's instructions, the fragmentation buffer was added, yet the fragmentation heating step was omitted. Again, qPCR quantification of mtDNA in pre- and post-WGA material was performed as described in section 2.3 on the same 96-well optical plate to minimize variation between quantifications.

2.8.3 Purification and Dilutions

WGA product from the QIAGEN® REPLI-g® Single Cell kit was suspected to cause unreliable qPCR quantification, as these samples resulted in high IPC values which indicated inhibition. On the other hand, the Sigma-Aldrich® GenomePlex® WGA2 kit showed lower IPC values than expected. Therefore, respectively 5 µl (eluted in 100 µl TE buffer) and 20 µl (eluted in 20 µl TE buffer) of product from these kits was purified using the Zymo® Clean & Concentrator-5™ kit (Zymo Research) as well as diluted 100-fold and 1000-fold with TE buffer prior to re-quantification with qPCR to test the efficiency of these methods in removing or diluting qPCR inhibitors.

In another purification experiment, Agencourt® AMPure® XP beads (Agencourt®) were used to purify the entire volume of WGA product from the QIAGEN® REPLI-g® Single Cell kit and the Sigma-Aldrich® GenomePlex® WGA2

kit. Subsequently, the Single Cell product was diluted 1000-fold and the GenomePlex® product was diluted 100-fold with sterile water. Then, both the AMPure® XP treated product and the diluted product were quantified with qPCR on the sample 96-well optical plate. The same purified product from the QIAGEN® REPLI-g® Single Cell kit was also diluted 100-fold with sterile water and quantified together with the 1000-fold dilution on the same 96-well optical plate with qPCR.

2.9 Long PCR on Hair Shaft Extract and WGA Material

The optimized extraction protocol used in this study may allow for the amplification of longer fragments from hair shaft extracts. Thus, a hair from a single donor was extracted as described in section 2.8.1 and eluted in 60 µl of elution buffer. The hair extract was then quantified with qPCR as described in section 2.3. From this extract, 30 µl (approximately 180,000 copies of mtDNA) was used as DNA input in an attempt to amplify the 9.1 kb LPCR segment as described in section 2.4.

In addition, an effort was made to amplify the 9.1 kb LPCR segment of a single donor from WGA product of the QIAGEN® REPLI-g® Single Cell kit and the Sigma-Aldrich® GenomePlex® WGA2 kit. This WGA product was purified with Agencourt® AMPure® XP beads as described in section 2.8.3. Again, LPCR conditions were followed as described in section 2.4, with mtDNA copy number input of 200,000, 500,000 and 1,000,000 copies of purified WGA product. Likewise, 200,000 and 500,000 copies of mtDNA of both post-WGA positive controls, as well as 10 µl of the post-WGA negative controls were amplified. The same copy numbers were targeted in an

amplification from WGA product that was purified and a thousand fold diluted with sterile water.

2.10 Multiplex Amplification and Next Generation Sequencing of WGA Product from Hair

E. Burnside performed two multiplexed PCR amplifications with two four- or five-plexes on the hair extracts and a subset of WGA product from each kit (Burnside *et al*, 2013). Primers were adapted for NGS from published primer sequences of the Applied Biosystems® mitoSEQr™ kit, and amplify mtDNA amplicons of approximately 400 - 700 bp in size. Table 11 shows the expected amplicon sizes for multiplex 1 (MP1) and multiplex 5 (MP5).

Table 11: Expected amplicon sizes for multiplex 1 (MP1) and multiplex 5 (MP5).

MP1 Amplicon Sizes (bp)	**MP5 Amplicon Sizes (bp)**
368	467
636	609
561	555
599	597
635	

Amplification was performed with Roche FastStart™ master mix, reaction conditions as per Table 12. The same conditions were used for 1 ng of a 9947A positive control.

Table 12: Reaction conditions for multiplex PCR with Roche FastStart™

Reagent	µl/reaction	Final concentration
Roche FastStart™ High Fidelity Reaction Buffer 10x, with 18 mM $MgCl_2$	2.5	1x, 1.8 mM $MgCl_2$
DMSO	2.5	10 %
dNTP Mix, 10 mM each	0.5	200 µM each
Roche FastStart™ High Fidelity Enzyme Blend	0.25	1.25 U
Primer mix	2.67	0.8 µM each
Sterile water	14.58	-
DNA template (various concentrations)	2	-
Total volume	**25**	**-**

Amplification was performed as per Table 13 on an Applied Biosystems® Veriti® 96-Well thermal cycler.

Table 13: Thermal cycling conditions for multiplex PCR with Roche FastStart™

95°C	2 min	
95°C	30 sec	36 cycles
55°C	30 sec	
72°C	1 min	
72°C	7 min	
4°C	hold	

In addition to performing the multiplex amplification, E. Burnside purified the PCR product with Agencourt® AMPure® XP beads, after which she quantified the resulting PCR product on the Agilent Technologies® 2100 Bioanalyzer® using the Agilent Technologies® DNA 1000 Kit™(Agilent Technologies 2006a). These quantifications were used in diluting the PCR product for NGS library preparation, by totaling the concentrations for each called peak and diluting this total concentration of DNA to 200 pg/µl.

2.10.1 Purification and Qubit® Quantifications

To determine DNA input for Illumina® Nextera® XT, a subset of WGA product from the QIAGEN® REPLI-g® Single Cell kit and the Sigma-Aldrich® GenomePlex® WGA2 kit was quantified with the Invitrogen™ Qubit® 2.0 Fluorometer using the Qubit® dsDNA HS (invitrogen 2010) and ssDNA (invitrogen 2011) assay kits. The dsDNA HS kit quantifies double-stranded DNA, however, the ssDNA kit is not specific to single-stranded DNA (Life Technologies). Typically, 1 - 5 µl of WGA product was used as input for quantification.

Subsequently, these samples were purified with a 1.8x ratio of Agencourt® AMPure® XP beads to WGA product (Agencourt®) to retain only single-stranded and double-stranded DNA of 100 bp and higher, and quantified again using both Qubit® kits.

In an attempt to remove the excess of single-stranded DNA, which is not fragmented by Illumina® Nextera® XT (Illumina® 2012), the WGA product was treated with USB® Exo-SAP-IT® (Affymetrix® 2011) and requantified with the Qubit® using both kits.

Because the REPLI-g® Single Cell and Sigma-Aldrich® GenomePlex® samples were highly concentrated in mtDNA (many millions of copies) it was unknown whether these would generate many more clusters on the Illumina® MiSeq™ flow cell as compared to the other samples that would be run simultaneously. In addition, since the samples were excessively handled, there was a high chance of cross-contamination. Therefore, it was decided not to sequence this product on this MiSeq™ flow cell, and instead sequence only the REPLI-g® Mini and Mitochondrial DNA products.

2.10.2 Illumina® Nextera® XT and Sequencing on Illumina® MiSeq™

Hair extracts processed with the REPLI-g® Mini and Mitochondrial DNA kits were quantified with the Invitrogen™ Qubit® 2.0 Fluorometer using the Qubit® dsDNA HS and ssDNA Assay kits to determine the total amount of dsDNA. WGA product was then diluted to 200 pg/µl of dsDNA based on these quantifications. In addition, the specific mtDNA copy number in these samples was quantified with qPCR. A second set of dilutions was based on molecular weight calculations based on these mtDNA quantifications.

Both sets of WGA dilutions were prepared for sequencing, with a subset of the duplicates from some donors pooled together in an attempt to ameliorate possible amplification bias. Combined with these samples were the multiplex amplified product and a negative control for the Nextera® XT process, which accounted for a total of 53 samples to be sequenced. These samples were each assigned a unique index combination and processed with Illumina® Nextera® XT as described before in section 2.6. All libraries were pooled with a 10% v/v spike-in of Illumina® PhiX Control v3 in the PAL. The PAL was then diluted 25-fold, to create the DAL. The DAL was sequenced on the Illumina® MiSeq™ in a 2x150 bp paired end v2 run. A sample sheet was designed and uploaded as described in section 2.6. Sequencing analysis was performed with Illumina® Sequence Analysis Viewer 1.8, Illumina® MiSeq™ Reporter 2.2, Integrative Genomics Viewer 2.3 and CLC bio® CLC Genomics Workbench 6.5. For the CLC Genomics Workbench analysis, parameters were as specified in Table 14.

Table 14: Analysis parameters for CLC bio® CLC Genomics Workbench 6.5. Default settings used unless otherwise specified in this table.

Trim Sequences:

- Quality trim
- 5' terminal nucleotides: 15 bases
- 3' terminal nucleotides: 10 bases
- Remove short reads (15 bases or shorter)

Map Reads to Reference:

- Using unmasked NC_012920
- Mismatch cost: 2
- Insertion cost: 3
- Deletion cost: 3
- Ignore non-specific matches

Quality-based Variant Detection:

- Minimum coverage: 100
- Minimum Variant Frequency: 0.5%
- Required & sufficient variant count: 10
- Require presence in both forward and reverse reads
- Ignore non-specific matches
- Minimum neighborhood quality: 30
- Minimum central quality: 30

With these parameters, each read is trimmed at both ends and short reads are filtered out. The remaining reads are also trimmed for quality, according to an algorithm developed by CLC bio®, after which they are aligned to the reference sequence (CLC bio). This strategy allows for higher quality mapping, since primer sequences and low-quality sequences are removed from the resulting data. For mapping, deletions and insertions are given a higher penalty than base mismatches, as few large gaps are expected in the mtGenome. If a read matches multiple regions of the reference, this read is ignored. For detection of variants from the rCRS, each position needs to be covered by a minimum number of 100 reads, and a variant needs to be detected in 10 reads or more

before it is called. In addition, the quality of the variant and the surrounding sequence should be Q30 or higher, and the variant should be seen in both forward and reverse reads. Again, non-specific matches are ignored.

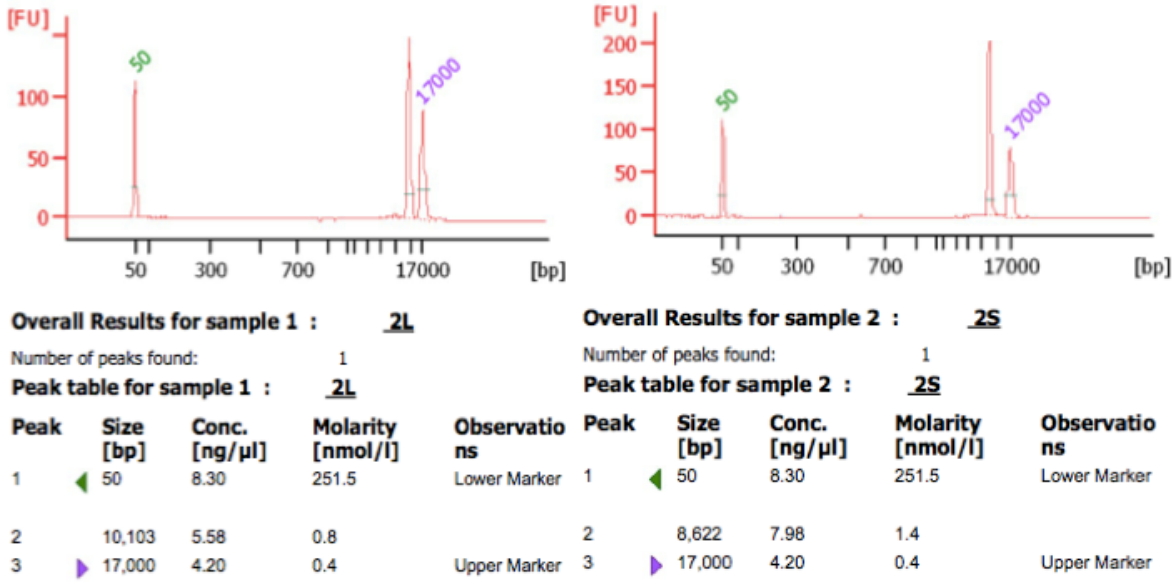
NGS data that were obtained for each WGA sample were compared to those derived from the LPCR samples on the MiSeq™ instrument for each donor (section 2.6). Positions that did not exhibit a common base in this comparison of treatments were designated as sequence differences.

CHAPTER 3: RESULTS

3.1 Long PCR Amplification

Reference data for eight donors was successfully obtained with the Applied Biosystems® mitoSEQr™ kit as described in section 2.2. A list of variants from the rCRS from each donor are listed in Appendix I.

To generate mtDNA template from reference samples for NGS, DNA was extracted from buccal swabs from eight donors and the mtDNA copy number was quantified with qPCR. Approximately 200,000 copies (Table 15) were used as input for LPCR. An example of LPCR product is shown in Figure 14.



Overall Results for sample 1 : 2L

Number of peaks found: 1

Peak table for sample 1 : 2L

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	50	8.30	251.5	Lower Marker
2	10,103	5.58	0.8	
3	17,000	4.20	0.4	Upper Marker

Overall Results for sample 2 : 2S

Number of peaks found: 1

Peak table for sample 2 : 2S

Peak	Size [bp]	Conc. [ng/µl]	Molarity [nmol/l]	Observations
1	50	8.30	251.5	Lower Marker
2	8,622	7.98	1.4	
3	17,000	4.20	0.4	Upper Marker

Figure 14: Quantification of long PCR product. Performed with the Agilent Technologies® 2100 Bioanalyzer® using the Agilent Technologies® DNA 12000 Kit™. Left: 11,170 bp amplicon. Right: 9,065 bp amplicon. Donor 002.

The average PCR product generated for each donor was approximately 6.6 ng/µl (Table 15). LPCR amplification failed a number of times on DNA extracted from cotton buccal swabs that were dried and stored at room temperature. Subsequently, successful amplification was observed when performed on fresh buccal swabs taken from donors. In consequence, DNA extraction was performed in three different batches of 2 or 3 donors at a time, with a separate reagent blank used for each batch. None of these showed LPCR amplification.

Table 15: Efficiency of long PCR amplification on buccal extracts. Average LPCR product is calculated as the average of the long and short amplicon per donor. A higher input for donor 003 was used because 1 ng of nuclear DNA was targeted for this amplification.

Donor	Copies of mtDNA in Buccal Swab Extract	LPCR Input (copies of mtDNA)	Average LPCR Product (ng/µl)
001	16,998,840,000	226000	5.35
002	62,612,828	208709	6.78
003	33,251,937	443359	10.52
006	18,411,570,000	246000	7.23
009	5,940,112,500	198000	5.30
015	1,037,101,905	230467	8.66
020	148,382,018	197843	7.34
021	54,837,990,000	183000	5.42

3.2 Long PCR Sequencing

After LPCR, the amplified product was processed with Illumina® Nextera® XT to generate sequencing libraries. Each amplicon was completely tagmented, into a broad peak of 100 - 400 bp visible on the Agilent Technologies® 2100 Bioanalyzer®. After normalization, a Qubit® quantification showed an average concentration of 250 pg/µl ssDNA for each sample.

Libraries were sequenced on the Illumina® MiSeq™ using v2 kit reagents in a 2x150 bp run. For this MiSeq™ run, 820K clusters/mm^2 were detected. Since Illumina® guidelines recommend a cluster density of 50 - 1300K/mm^2, with an optimum of 800K/mm^2, this run was accepted (Illumina® 2013a). In addition, 84.28% of the quality scores were Q30 or higher, and 90.19% of all clusters Passed Filter (PF), a quality-filtering step.

All NGS data was analyzed with MiSeq™ Reporter (MSR) 2.2. Whole mtGenome data was obtained for each donor, as seen in Figure 15. The Sanger sequences (Appendix I) were compared to the NGS sequences (Appendix II) for each donor, and positions that did not exhibit a common base between the two compared sequences were designated as sequence differences. It should be kept in mind that due to the low resolution of Sanger sequencing, low-level mixtures or mixed positions may not be detected in the Sanger analysis. Therefore, some of the mixed positions detected in the comparison may be due to the distinctive levels in resolution obtained with the two methods.

It should be noted that bioinformatics software packages have known limitations with base calling in sequences that contain small insertions and deletions (indels). As the reads containing indels can independently be mapped to a reference sequence, this may result in multiple variant calls for a single indel (Albers et al. 2011). Therefore, misalignments and small indels in NGS data are omitted from the analysis results in this study.

In MSR, lower coverage is observed at the distal ends of the mtGenome. This may be due to the software's inability to recognize the reference sequence as circular, which causes a lowering in coverage in these regions.

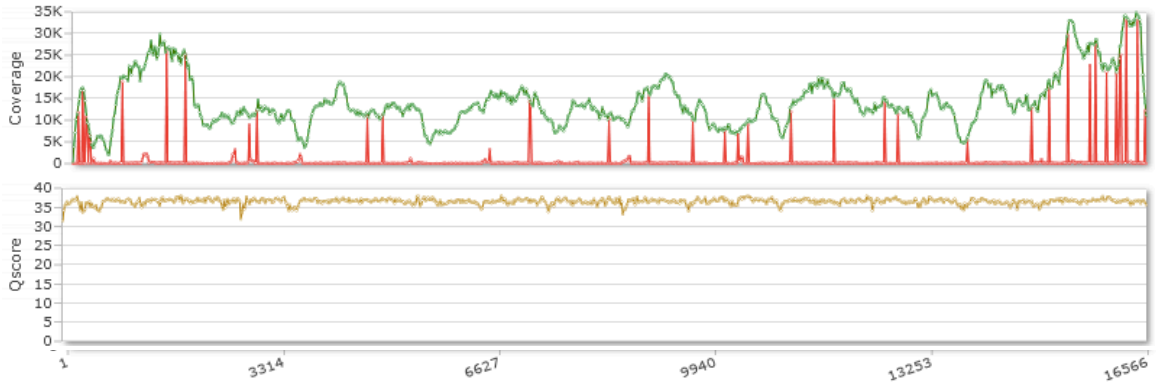


Figure 15: Whole mtGenome coverage graph from MiSeq™ Reporter for donor 002. Top: read coverage across the genome. Bottom: Quality scores.

The average whole mtGenome coverage of sequencing data for all donors was 13072 reads in the MSR analysis (Table 16). The NGS data revealed 11 to 41 variants from the rCRS outside of the HV regions, with an average of 26. The median fragment length across all donors was 265 bp, which is consistent with the Agilent Technologies® 2100 Bioanalyzer® size distributions of the Illumina® Nextera® XT libraries.

Table 16: Variants from the rCRS, coverage and fragment lengths in whole mtGenome NGS data. Analysis performed with MSR.

Donor	Variants Outside of HV Regions	Median Coverage	Median Fragment Length (bp)
001	23	6391	273
002	28	13706.1	262
003	27	13506.5	258
006	11	17213.4	266
009	12	16077.9	253
015	41	15573.8	254
020	31	12596.8	278
021	36	9506.1	276

In Table 17, an example of NGS data from one donor in this MiSeq™ run is shown. Data for the other donors can be found in Appendix II.

Table 17: Variants from the rCRS in NGS and Sanger sequencing data from donor 002. Data was analyzed with MSR. Yellow: common base between Sanger and NGS analysis; Pink: low-level mixed position; Blue: low-level mixed position in homopolymer region.

Sanger			Illumina® MiSeq™					Sanger			Illumina® MiSeq™				
Pos	rCRS	Var	Pos	Type	Call	Freq	Depth	Pos	rCRS	Var	Pos	Type	Call	Freq	Depth
73	A	G	73	SNP	A->AG	100	11289	8,860	A	G	8860	SNP	A->AG	100	15077
152	T	C	152	SNP	T->TC	100	16694	9,548	G	A	9548	SNP	G->GA	100	9238
199	T	C	199	SNP	T->TC	100	10632	10,034	T	C	10034	SNP	T->TC	100	7260
204	T	C	204	SNP	T->TC	100	9558	10,238	T	C	10238	SNP	T->TC	100	6587
207	G	A	207	SNP	G->GA	100	9341	10,398	A	G	10398	SNP	A->AG	100	8828
250	T	C	250	SNP	T->TC	100	5959	11,065	A	G	11065	SNP	A->AG	100	11494
263	A	G	263	SNP	A->AG	100	4512	11,719	G	A	11719	SNP	G->GA	100	14204
309.1	:	C	302	Indel	-/C	91	1755	12,501	G	A	12501	SNP	G->GA	100	8863
315.1	:	C	310	Indel	-/C	100	2043	12,705	C	T	12705	SNP	C->CT	100	11234
573.1	:	C	567	Indel	---/CCC	49	1781	13,780	A	G	13780	SNP	A->AG	100	4520
750	A	G	750	SNP	A->AG	100	18207	14,766	C	T	14766	SNP	C->CT	100	12300
1,438	A	G	1438	SNP	A->AG	100	25567	15,043	G	A	15043	SNP	G->GA	100	16826
1,719	G	A	1719	SNP	G->GA	100	24450	15,326	A	G	15326	SNP	A->AG	100	28723
2,706	A	G	2706	SNP	A->AG	100	6461	15,673	A	G	15673	SNP	A->AG	83	26461
2,835	C	A	2835	SNP	C->CA	100	11764	15,758	A	G	15758	SNP	A->AG	100	26543
3,107	N	:	3106	Indel	N/-	94	10710	15,924	A	G	15924	SNP	A->AG	100	20390
4,529	A	T	4529	SNP	A->AT	100	10163	16,074	A	G	16074	SNP	A->AG	100	20066
4,769	A	G	4769	SNP	A->AG	100	11051	16,129	G	A	16129	SNP	G->GA	99	23467
7,028	C	T	7028	SNP	C->CT	99	13846	16,145	G	A	16145	SNP	G->GA	100	24327
7,055	A	T	7055	SNP	A->AT	100	12759	16,223	C	T	16223	SNP	C->CT	99	31446
8,251	G	A	8251	SNP	G->GA	100	9854	16,391	G	A	16391	SNP	G->GA	100	31781
8,843	T	T	*8843	SNP	T->TC	2	16098	16,519	T	C	16519	SNP	T->TC	100	11915

NGS has enhanced capability to detect sequence mixtures. For example in donor 001, an approximately 8% known low-level mixed position was detected at position 16,093 (Appendix II). In the data set from donor 002 in Table 17, an "A" was observed at approximately 17% at position 15,673. Upon revisiting the Sanger sequence electropherograms for this donor, a mixed position was indeed present at 15.673 and the Sanger data was amended to include this finding (Figure 16).

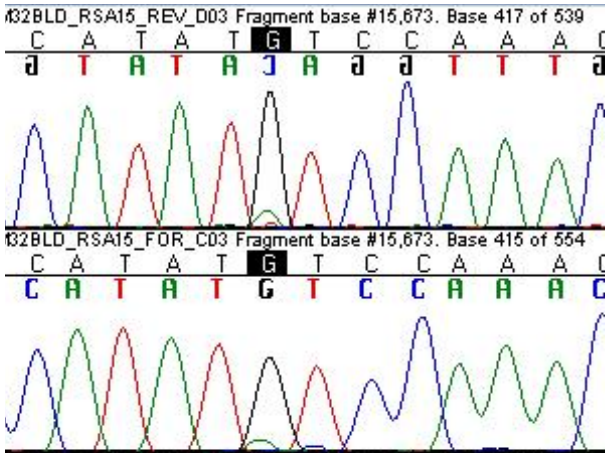


Figure 16: Mixed base at position 15,673 in donor 002. This mixed position was noted in the Sanger sequence data after the NGS analysis.

Another category of low-level mixed positions was found in the MSR data, indicated in blue in Table 17. These occur around homopolymer regions (a contiguous occurrence of the same nucleotide). These mixed positions are not a result of primer binding mutations, which occur due to the incorporation of an amplification primer that contains base mismatches compared to the template but is still able to amplify this template. The observed mixed positions also do not cluster in specific reads, as would be seen with nuclear insertions of mitochondrial DNA (NUMTs) or arising from external contamination (Blanchard and Schmidt 1996). One particular deletion found at position 12,417 (eight consecutive adenines) is seen at a frequency of approximately 4% in MSR data from all donors. Although this deletion was ignored in the analysis, it should be noted that it was also seen in MSR analyses of other MiSeq™ runs. These may be caused by a mixed position or could be an artifact of alignment as explained previously. Further evaluation of these deletions is warranted.

3.3 Sequencing of NIST Standards

The DNA sequencing standards described in section 2.5 that were obtained from the National Institute of Standards and Technology were sequenced to evaluate their use as sequencing controls. SRM 2392 consisted of three human DNA standards, which were two DNA extracts and a cloned HVI region, whereas SRM 2394 consisted of ten different mixtures of two amplicons with a single base mismatch at the same position. As the NIST standards were sequenced in the same MiSeq™ run as the long PCR amplicons, the same run statistics as stated in section 3.2 apply.

None of the NIST standards were PCR amplified prior to sequencing. In consequence, all three SRM 2392 standards show low coverage, although ssDNA quantification of the sample post-normalization did show ssDNA quantities similar to the LPCR samples. Approximately 40 reads (40x) were observed across the entire mtGenome for both extracts, however these also showed reads mapping to the nuclear genome. A median of 950x coverage was detected for the cloned section of HV1. Adequate read depth is important to reliably detect low-level mixtures. As a result, it was not possible to call all variants from the rCRS for these samples.

The SRM 2394 standards were analyzed with MSR. Depth of coverage across the amplicons varied. All showed high coverage, with an average depth of approximately 412,000 reads at the position of the mixed base (Table 18). The median coverage across the entire amplicon was often higher, up to twice as high on one occasion. One of the samples was sequenced at a lower depth than the others, but it was possible to detect a low-level mixture. The reported frequency of bases was very close to the expected frequencies reported by NIST.

Table 18: Evaluation of NGS data accuracy from NIST Mixture Standards. Expected and called mixture frequencies of base position 6,317 in all ten standards. Data derived from MSR.

Sample ID	Base expected	Freq. (%)	Base called	Freq. (%)	Depth
NISTMix1	T	100	T	99	274063
NISTMix2	C	100	C	100	394144
NISTMix3	T/C	50	T/C	52	468301
NISTMix4	T/C	40	T/C	41	467779
NISTMix5	T/C	30	T/C	31	566795
NISTMix6	T/C	20	T/C	22	381094
NISTMix7	T/C	10	T/C	11	480589
NISTMix8	T/C	5	T/C	6	319786
NISTMix9	T/C	2.5	T/C	4	39130
NISTMix10	T/C	1.0	T/C	2	727996

3.4 Whole Genome Amplification on Buccal Swabs

Whole Genome Amplification studies were initially performed on buccal swab extracts to assess the efficacy in augmenting mtDNA in a robust sample type. This was determined by the "fold increase" of mtDNA in the WGA material: the increase in mtDNA copy number as compared to the input. In the first experiment, a dilution series was based on nDNA quantification. WGA with all four kits mentioned in section 1.5.2 was performed on these dilutions, prepared from the extracted DNA of one donor (Table 19/Figure 17).

Table 19: First experiment: WGA performed on diluted buccal extract from a single donor. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Lowest, average and highest fold increase calculated for all donor samples, controls excluded. Italics: copies not observed in all triplicate quantifications.

Input		**Fold Increase**			
nDNA	*mtDNA copies/ul	Q®R® Mitochondrial DNA	Q®R® Mini	S-A® GenomePlex® WGA2	Q®R® Single Cell
100 pg/µl	13,282	185	945	830	31,751
50 pg/µl	7,441	82	1,262	853	30,089
25 pg/µl	3,265	2,409	3,073	891	44,207
12.5 pg/µl	1,758	155	2,984	892	77,338
6.25 pg/µl	740	1,436	5,856	946	303,939
3.13 pg/µl	380	3	8,300	1,024	339,893
1.56 pg/µl	165	198	29,581	1,529	287,843
785 fg/µl	79	746	20,108	785	1,163,055
392.5 fg/µl	64	266	8,520	1,188	681,678
196.25 fg/µl	31	16	466	1,021	7
QIAGEN® Pos	2,647,627	5,499	87	1,839	621
WGA Neg	0	0	0	0	0
RB	0	*5*	0	*41*	*5*
Lowest		3	466	785	7
Average		**550**	**8,109**	**996**	**295,980**
Highest		2,409	29,581	1,529	1,163,055

*5 µl used for Sigma-Aldrich® GenomePlex® WGA2 Kit and QIAGEN® REPLI-g®, Mitochondrial DNA and Mini kits, 2.5 µl used for QIAGEN® REPLI-g® Single Cell kit

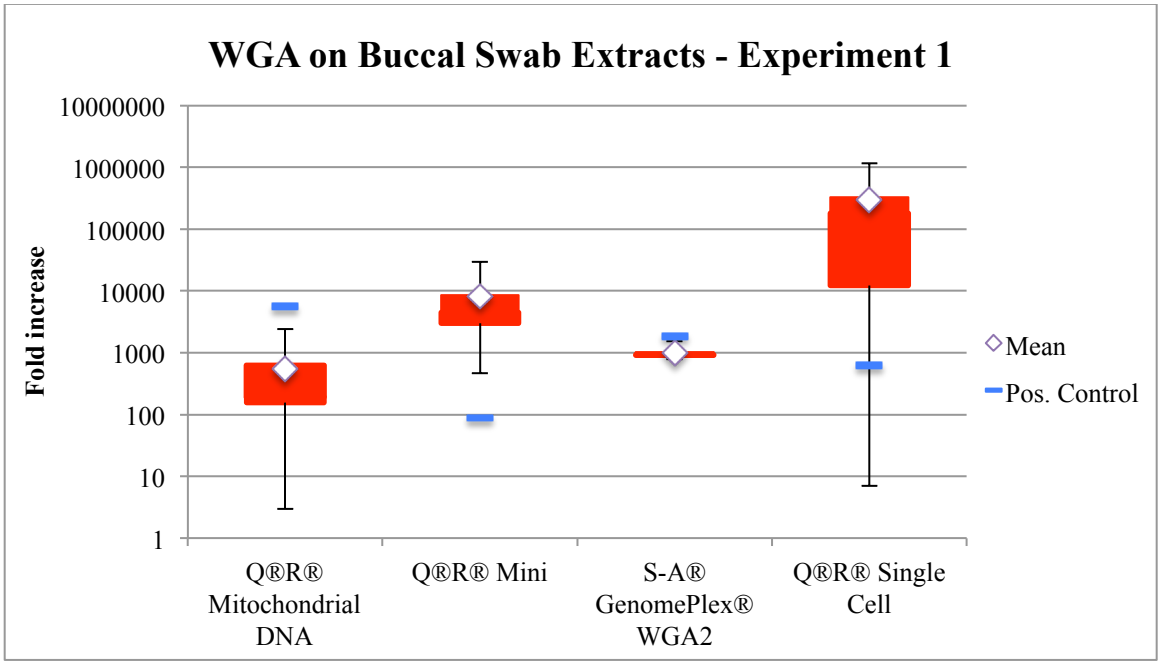


Figure 17: First experiment: WGA performed on diluted buccal extract from a single donor. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Box: lower and upper quartile limits. Whiskers: lowest and highest fold increases.

It should be noted that these experiments with buccal swab extractions were performed before it was discovered that material in WGA product from the QIAGEN® REPLI-g® Single Cell kit interferes with the qPCR quantification (see section 3.6). Therefore, the reported values for the Single Cell kit may underestimate the actual mtDNA concentration in these samples.

In the second experiment, another dilution series different from the first experiment was prepared from the extracted DNA of two donors and WGA was performed on these dilutions (Table 20/Figure 18).

Table 20: Second experiment: WGA performed on diluted buccal extracts from two donors. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Lowest, average and highest fold increase calculated for all donor samples, controls excluded. Italics: copies not observed in all triplicate quantifications.

Input		**Fold Increase**				**Fold Increase**	
nDNA	*mtDNA copies/ul	S-A® GenomePlex® WGA2	Q®R® Mini Trial 1	Q®R® Mini Trial 2	*mtDNA copies/ul	Q®R® Mitochondrial DNA	Q®R® Single Cell
002 100 pg/µl	24,696	395	27	105	44,007	180	30,596
002 3.13 pg/µl	816	552	19	217	1,388	310	294,612
002 196.25 fg/µl	51	1,003	85	5	86	22	228,322
020 100 pg/µl	12,084	541	60	218	20,191	609	15,087
020 3.13 pg/µl	379	581	59	293	510	17	222,015
020 196.25 fg/µl	29	193	103	125	29	3	14,791
RB	*4*	0	0	0	*4*	0	0
QIAGEN® Pos	118,617	981	96	147	99,345	1,707	8,848
WGA Neg	0	0	0	0	0	0	0
Lowest		193	19	5		3	14,791
Average		**544**	**59**	**160**		**190**	**134,237**
Highest		1,003	103	293		609	294,612

*5 µl used for Sigma-Aldrich® GenomePlex® WGA2 and QIAGEN® REPLI-g® Mini kits, 2.5 µl used for QIAGEN® REPLI-g® Single Cell and Mitochondrial DNA kits

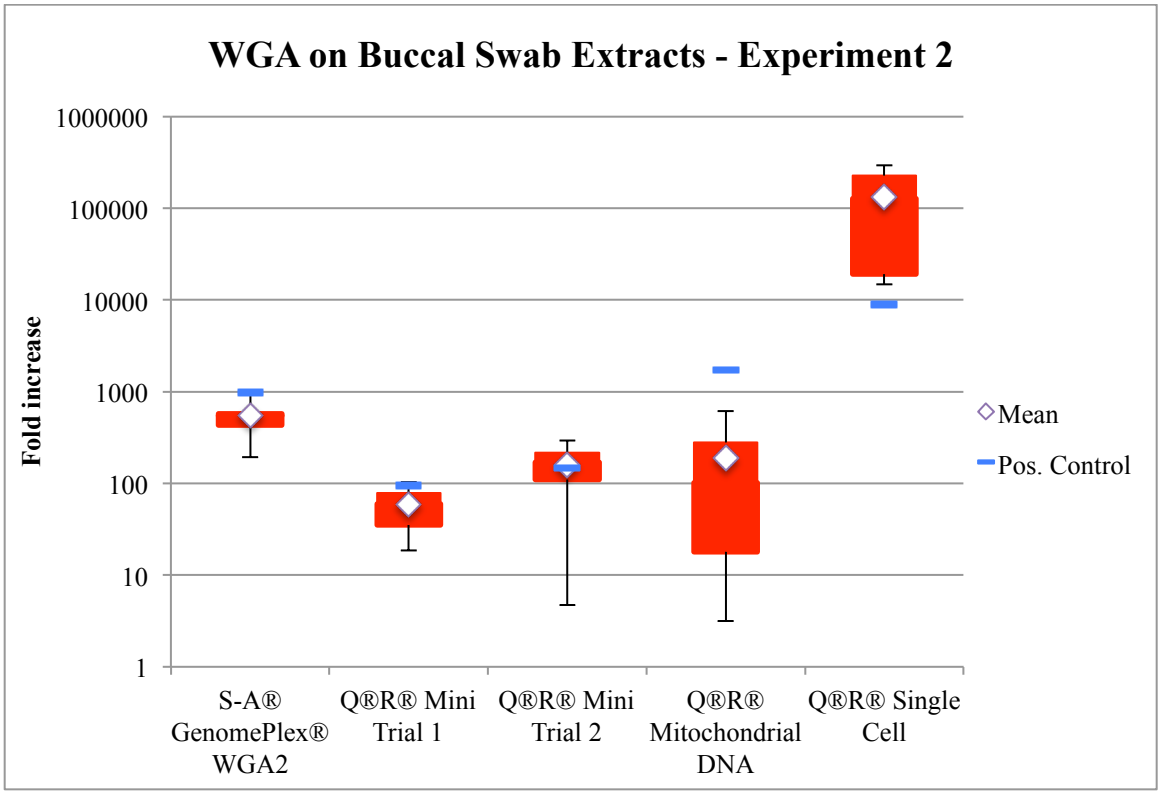


Figure 18: Second experiment: WGA performed on diluted buccal extracts from two donors. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Box: lower and upper quartile limits. Whiskers: lowest and highest fold increases.

In the third experiment, a dilution series was again prepared from the extracted DNA of two donors and WGA was performed on these dilutions, however this time the dilutions were based on a specific mtDNA copy number, not an nDNA concentration (Table 21/Figure 19-20).

Table 21: Third experiment: WGA performed on diluted buccal extracts from two donors. Q®R®: QIAGEN® REPLI-g®. Lowest, average and highest fold increase calculated for all donor samples, controls excluded.

Input		Fold Increase	
Sample ID	*mtDNA copies/µl	Q®R® Mini	Q®R® Mitochondrial DNA
002	20250	565	27,598
002	6750	1,013	12,348
002	2250	1,157	26,029
002	750	289	403
002	250	11,971	44,278
020	20250	863	7,789
020	6750	1,135	24,679
020	2250	2,169	4,288
020	750	3,413	188,171
020	250	4,748	14,847
RB	0	0	0
QIAGEN® Pos	69,560	256	4,949
HL60 Pos	18,955	480	1,320,342
WGA Neg	0	0	0
Lowest		289	403
Average		**2,732**	**35,043**
Highest		11,971	188,171

*Used 5 µl input for both kits

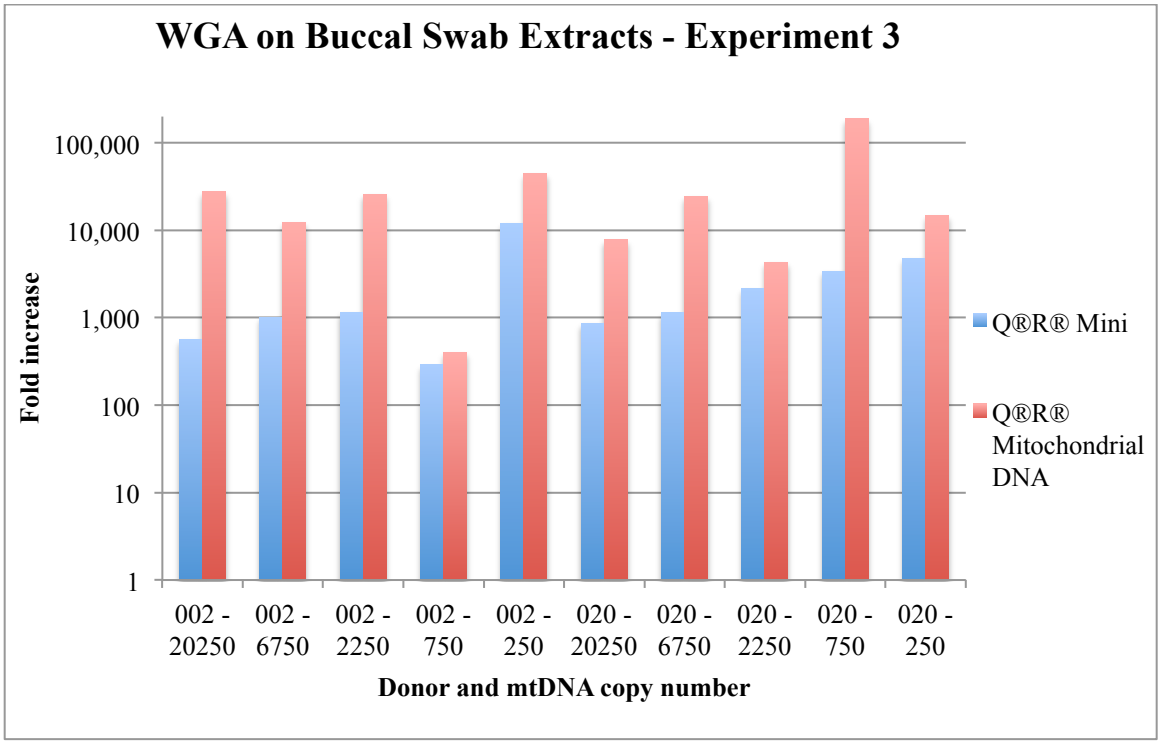


Figure 19: Third experiment: bar chart of WGA performed on diluted buccal extracts from two donors. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®.

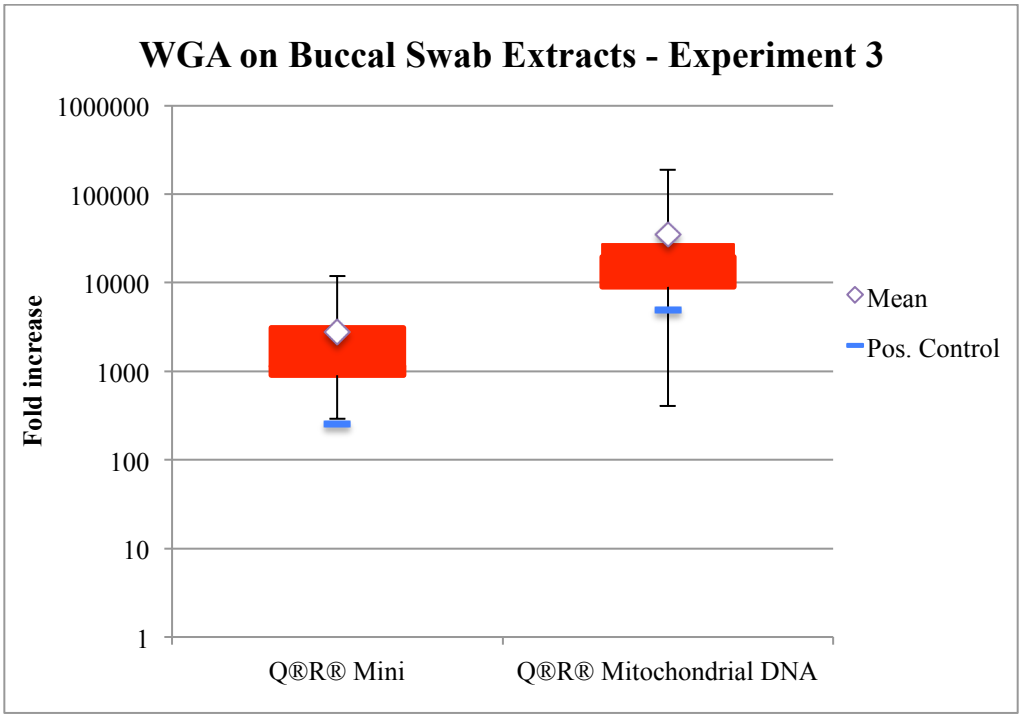


Figure 20: Third experiment: box plot of WGA performed on diluted buccal extracts from two donors. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Box: lower and upper quartile limits. Whiskers: lowest and highest fold increases.

It should be noted that these results are preliminary since limited data is available and there was a wide range of values obtained from many of the replicates. Fold increases from experiment to experiment were inconsistent for all kits except the Sigma-Aldrich® GenomePlex® WGA2 kit. Fold increases are not consistent with copy number input for any of the WGA kits. This inconsistency may be due to low DNA input into the WGA reactions: many WGA manuals recommend 1-10 ng of DNA template as input. Differences may also lie in pipetting accuracy: small changes in input may cause noticeable differences in fold increases. Lastly, differences may be due to chance. A few more primer binding events in the initial stages of the MDA process could facilitate a higher degree of amplification in total.

In these few preliminary experiments, it is evident that the QIAGEN® REPLI-g®

Single Cell kit resulted in a high fold increase in mtDNA copy number. The QIAGEN® REPLI-g® Mitochondrial DNA and Mini kits as well as the Sigma-Aldrich® GenomePlex® WGA2 kit showed lower fold increases. These results are concordant with previous studies, which show inconsistent fold increases in mtDNA in pristine samples with lower mtDNA concentrations (Maragh et al. 2008).

In the first experiment, the QIAGEN® REPLI-g® Mini kit showed a higher average fold increase than in the second or third experiment. Therefore, during the second experiment a second trial was done with new reagents, however the increases did not return to the levels seen in the first experiment.

3.5 Whole Genome Amplification on Hair Shaft Extract

After experiments with buccal extracts, Whole Genome Amplification was performed on hair shaft extracts to assess the efficacy in augmenting the mtDNA copy number in challenging sample types. DNA from a single hair shaft was extracted from each of three donors. After each DNA extraction, WGA was performed in duplicate for each hair shaft extract, after which the increase in mtDNA copy number was assessed with qPCR. This experiment was replicated three times, using freshly extracted DNA from hair shafts each time (Tables 22, 23 and 24/Figure 21).

Table 22: WGA performed on hair shaft extract from three donors, first experiment. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. Lowest, average and highest fold increase calculated for all donor samples, controls excluded. Italics: copies not observed in all triplicate quantifications.

Input		**Fold Increase**		
Sample ID	mtDNA copies/µl	Q®R® Mitochondrial DNA	S-A® GenomePlex® WGA2	Q®R® Mini
002-1	28,462	2	429	3
002-2	28,462	3	484	2
009-1	7,942	2	422	2
009-2	7,942	2	308	2
020-1	19,834	3	627	12
020-2	19,834	3	744	54
HL60 Pos	78,967	233,801	921	745
RB	0	0	*20*	0
WGA Neg	0	0	*33*	*7*
Lowest		2	308	2
Average		**2**	**502**	**13**
Highest		3	744	54

Table 23: WGA performed on hair shaft extract from three donors, second experiment. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. NF: No Fragmentation step. Lowest, average and highest fold increase calculated for all donor samples, controls excluded. Italics: copies not observed in all triplicate quantifications.

Input		**Fold Increase**			
Sample ID	mtDNA copies/µl	Q®R® Mini	S-A® GenomePlex® WGA2	Q®R® Mitochondrial DNA	S-A® GenomePlex® WGA2 NF
002-1	8,353	1	179	1	328
002-2	8,353	1	176	1	-
009-1	10,257	1	196	1	504
009-2	10,257	1	340	1	-
020-1	10,637	2	389	1	793
020-2	10,637	1	476	2	-
HL60 Pos	49,613	340	927	99,213	1,125
RB	0	7	*23*	*2*	*9*
WGA Neg	0	4,368	0	0	*5*
Lowest		1	176	1	328
Average		**1**	**293**	**1**	**542**
Highest		2	476	2	793

Table 24: WGA performed on hair shaft extract from three donors, third experiment. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. NF: No Fragmentation step. Lowest, average and highest fold increase calculated for all donor samples, controls excluded.

Input		**Fold increase**	
Sample ID	mtDNA copies/µl	Q®R® Single Cell	S-A® GenomePlex® WGA2 NF
002-1	5,965	17,373	271
002-2	5,965	144,679	360
020-1	10,030	1,176	638
020-2	10,030	109,052	884
HL60 Pos	39925	31,745	823
RB	0	0	0
WGA Neg	0	0	0
Lowest		5	0
Average		**45,384**	**359**
Highest		144,679	884

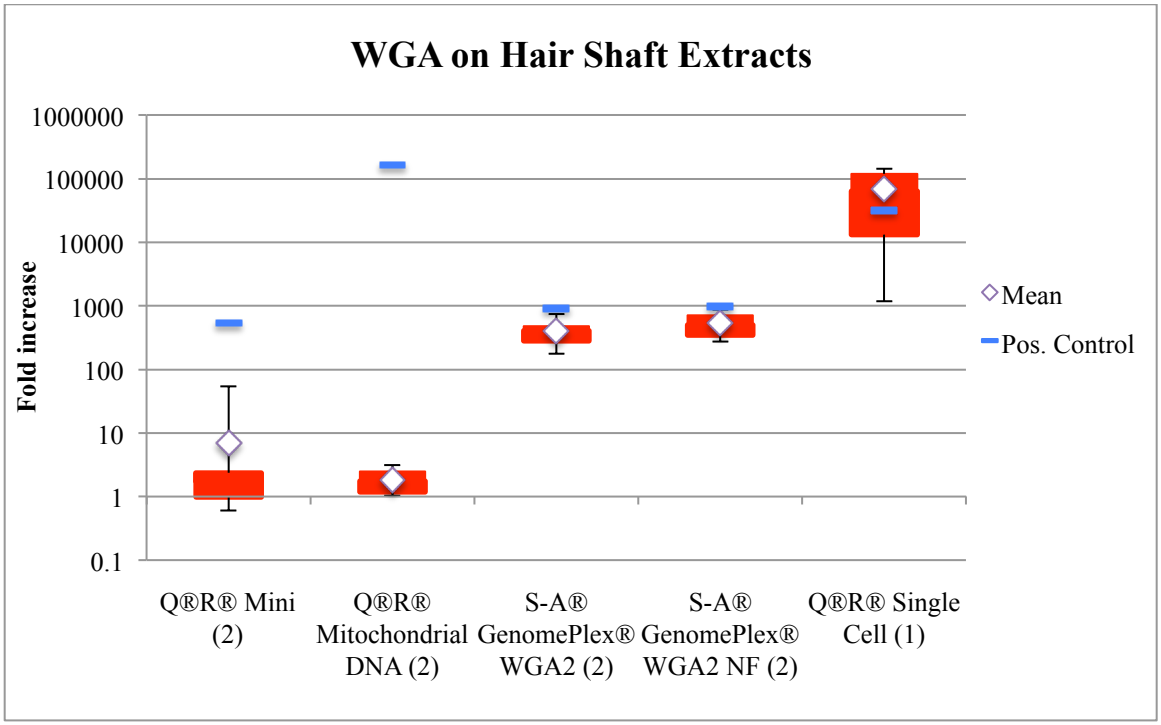


Figure 21: All WGA experiments on hair shaft combined. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. NF: No Fragmentation step. Number of replicates noted in parentheses behind method name. Box: lower and upper quartile limits. Whiskers: lowest and highest fold increases.

Similar to the experiments with DNA from buccal swabs, WGA on hair shaft DNA extracts also results in inconsistent increases in mtDNA copy number for all kits except the Sigma-Aldrich® GenomePlex® WGA2 kit. In this study, the QIAGEN® REPLI-g® Single Cell kit shows higher fold increases in mtDNA copy number from hair shaft extracts than is exhibited by the QIAGEN® REPLI-g® Mitochondrial DNA and Mini kits and the Sigma-Aldrich® GenomePlex® WGA2 kit. Again, it should be noted that these results are preliminary, since a wide range of values was obtained from many of the replicates.

In the third batch of extractions, the hair shaft extract from donor 009 yielded very little mtDNA, most likely due to an error during extraction. This extract showed low fold increases after WGA with both kits, and is therefore excluded from these results.

3.6 Purification of WGA Product

3.6.1 Obtaining Accurate qPCR Quantification Values

The IPC assay used in qPCR quantifications is used to verify whether the reagents and the instrument are working properly and to detect the presence of PCR inhibitors. Usually, IPC values exhibit a mean C_T of approximately 28 on the ABI PRISM® 7000 Sequence Detection System and a C_T of 25 on the Applied Biosystems® 7500 Real-Time PCR System. These C_T values normally increase as the DNA concentration in samples or standards increases due to competition for reagents; elevation becomes noticeable at mtDNA copy numbers of 10,000 - 100,000 and up (Kavlick et al. 2011). Upon qPCR

quantification of mtDNA in the WGA product it was evident that for the QIAGEN® REPLI-g® Single Cell kit, Internal Positive Control (IPC) C_T values were elevated in the WGA product (C_T = undetermined for most samples on both the 7000 and 7500 instrument). However, the IPC C_Ts for the Single Cell kit were also elevated in the post-WGA negative controls, which did not contain any mtDNA. Although Single Cell samples from hair shaft contained millions of copies of mtDNA, which could affect the assay as mentioned before, the negative control indicated that reagents in the Single Cell kit also attributed to elevated qPCR IPC C_T values. The effects of WGA kits on qPCR need to be better understood in order to obtain accurate qPCR quantification results.

A similar occurrence was seen with the Sigma-Aldrich® GenomePlex® WGA2 kit, yet IPC values were often at the same level or lower than the negative controls (C_T = 24 for the 7500 instrument, and 28 for the 7000) even though changes in IPC C_T values should have been observed as these samples contained high concentrations of mtDNA. This indicated that reagents from the GenomePlex® kit were possibly influencing the qPCR assay as well.

In an effort to further study the qPCR effects in these samples, a subset of samples from both kits were purified with Agencourt® AMPure® XP magnetic beads or the Zymo® Clean & Concentrator-5™ column purification kit. Upon requantification, IPC values did change, however this was consistent with the high mtDNA concentrations in these samples, which can elevate IPC values due to reagent depletion as mentioned previously. After diluting the products of both kits (in some cases purified, in some cases not) a 100-fold and 1000-fold with sterile water, results did no longer indicate the presence of inhibitors.

It should be noted there was no change between mtDNA copy number results for the GenomePlex® samples before or after purification/dilution. However, these treatments were necessary to facilitate successful PCR amplification in these samples (section 3.7). The Single Cell kit results after purification/dilution showed that these samples contained more mtDNA copies than was initially quantified. These results suggest that WGA product arising from the use of these kits may require purification and dilution prior to the downstream applications.

As stated previously, the inhibitory effect in Single Cell kit reagents was not detected during the WGA experiments with buccal swab extracts. Therefore, fold increases in mtDNA for samples from the Single Cell kit may be underestimated in these experiments.

3.6.2 DNA Purification Prior to Library Preparation

To determine the optimal DNA input for Illumina® Nextera® XT processing, single-stranded and double-stranded DNA (ssDNA and dsDNA) were quantified with the Qubit® 2.0 using both the ssDNA and dsDNA quantification kits. The ssDNA quantification was performed because ssDNA is not a substrate for Illumina® Nextera® XT, and it was unknown whether excess amounts of ssDNA would interfere with the library preparation method. It should be noted that the dsDNA kit quantifies dsDNA specifically, yet the ssDNA kit quantifies both ssDNA as well as dsDNA and RNA in a sample (Life Technologies). Therefore, both dsDNA and ssDNA quantifications need to be performed and the quantification value for dsDNA should be subtracted from the

ssDNA quantification value if the ssDNA concentration is sought. In addition, both kits are susceptible to varying degrees of signal change due to contaminants such as salts and organic solvents.

To determine the total dsDNA and ssDNA content of WGA products, an aliquot of product from the QIAGEN® REPLI-g® Single Cell kit and the Sigma-Aldrich® GenomePlex® WGA2 kit were quantified with both Qubit® kits. Both sample types initially showed a high ratio of ssDNA to dsDNA. Samples were purified with Agencourt® AMPure® XP beads and requantified. This resulted in a higher concentration of dsDNA in Single Cell kit product than was reported previously, which indicates that components in this kit or the WGA product influence the Qubit® quantification. However, the concentration of dsDNA and ssDNA in the GenomePlex® product was lowered after purification, which is consistent with the removal of primers and primer complexes by the magnetic beads.

Since ssDNA is not a substrate for Illumina® Nextera® XT, an attempt was made to remove ssDNA from the bead-purified WGA product using USB® ExoSAP-IT®, after which the product was requantified. High amounts of ssDNA were present in the Single Cell product, which was not completely removed after two purification steps with USB® ExoSAP-IT®, as assessed by Qubit® quantification.

3.7 Multiplex Amplification of Hair Shaft Extract and WGA Material

Two different multiplex PCRs, each containing 4 or 5 different primer sets that target different portions of the mtGenome, were performed on a subset of hair shaft

extracts and WGA product from all four kits. Multiplex amplification was successful on unpurified WGA product of the QIAGEN® REPLI-g® Mitochondrial DNA kit (Figure 22) and Mini kit (data not shown). Due to the low mtDNA augmentation in these samples it cannot be confirmed whether it was WGA product that was PCR amplified or whether or not this signal originated from PCR amplification of the initial DNA extract. However, these results do indicate that the Mini and Mitochondrial DNA kits likely do not contain reagents that are inhibitory to multiplex PCR amplification.

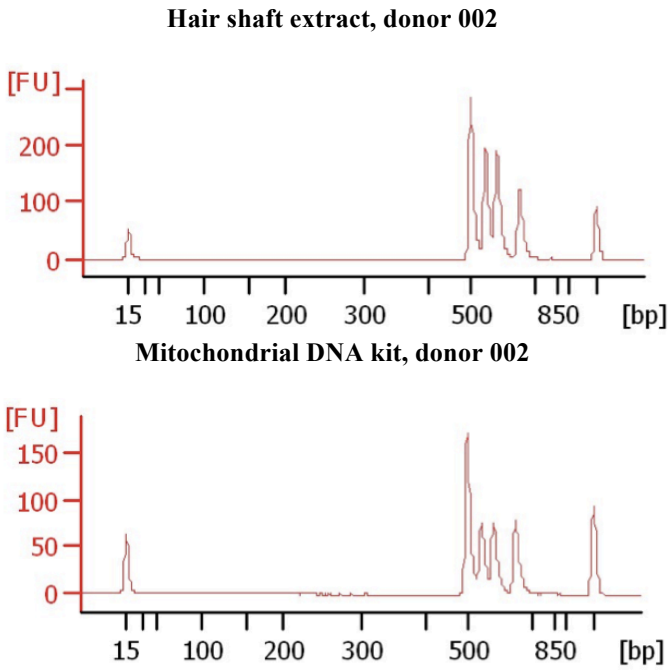


Figure 22: Multiplex PCR performed on hair extract and WGA material of donor 002. Material from the first hair shaft experiment, amplified with MP5 primers.

Top: donor 002 extract, 26,554 copies of mtDNA input. Total concentration of PCR product 24.06 ng/µl.

Bottom: corresponding WGA product from the QIAGEN® REPLI-g® Mitochondrial DNA kit, 4,182 copies input. Total concentration of PCR product 12.39 ng/µl.

No multiplex PCR amplification was obtained with unpurified WGA product from the GenomePlex® kit, using inputs of 90,000 to a 263,000 copies of mtDNA. Upon purification of the WGA product with the Zymo® Clean & Concentrator-5™ kit or a 100-fold to a 1000-dilution with sterile water, amplification was successful and generated sufficient product for NGS library preparation (Figure 23). This same pattern was observed for the Single Cell kit (data not shown).

In a 1000-fold diluted WGA sample only limited quantities of mtDNA from the original hair shaft extract remain. Therefore, in some cases the WGA product itself may be the supporting template for the observed successful multiplexed PCR amplification. However, these hair shaft extracts should have been diluted 1000-fold prior to multiplex amplification and sequencing, to determine whether or not the diluted extracts could support multiplexed PCR amplification at such a dilution factor. Thus, further studies are required to confirm these findings.

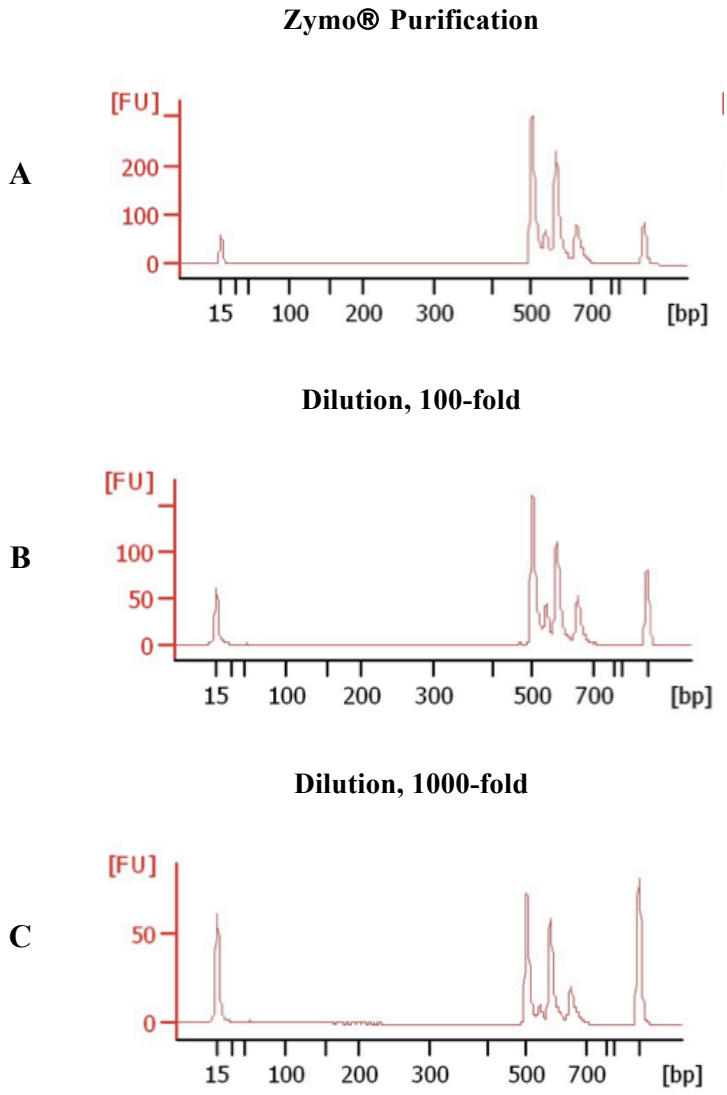


Figure 23: Multiplex PCR is successful after purifying or diluting WGA product. Product from Sigma-Aldrich® GenomePlex® WGA2 kit, donor 020. Material from the second hair shaft experiment, amplified with MP5 primers.

A: WGA product purified with Zymo® Clean & Concentrator-5™ kit. Total concentration of PCR product 25.41 ng/µl.

B: WGA product diluted 100-fold. Total concentration of PCR product 13.32 ng/µl.

C: WGA product diluted 1000-fold. Total concentration of PCR product 5.56 ng/µl.

3.8 LPCR on Hair Shaft Extract and WGA Material

An attempt to amplify the 9.1 kb LPCR fragment from a hair extract was not successful with 180,000 copies of mtDNA input (30 µl input). Whole Genome Amplification was performed with both the QIAGEN® REPLI-g® Single Cell kit, as well as the Sigma-Aldrich® GenomePlex WGA2 kit (fragmentation step omitted). Amplification of the 9.1 kb LPCR segment was only successful on an HL60 positive control that was amplified with the Single Cell kit and diluted to 200,000 and 500,000 copies of mtDNA input. These samples resulted in 1.39 ng/µl and 13.17 ng/µl of LPCR product, respectively, although non-specific amplification of smaller fragments was observed. LPCR amplification was minimally successful (0.35 ng/µl) with 500,000 copies of mtDNA from the same HL60 positive control used as input for the GenomePlex® kit. None of the WGA product from hair shaft extracts was successfully amplified with LPCR.

It should be noted that using 200,000 copies of mtDNA from a pristine positive control as input, product from both WGA kits yielded LPCR fragments at a concentration that was lower than usually observed at this level of input (approximately 5 - 10 ng/µl). These results indicate that these WGA methods likely shorten the DNA template during the amplification reaction. As mentioned in section 1.5.2, it is unlikely for MDA random hexamer primers to consistently anneal at the distal ends of a fragment. Thus, if a hexamer e.g. anneals in the middle of a fragment, only half of that fragment is amplified. This causes shortening of the DNA template. The GenomePlex® method incorporates a fragmentation step, and although this step may be omitted (section 2.8.2), it is possible

that the subsequent PCR amplification steps are not designed for amplification of longer, non-fragmented, template DNA strands.

3.9 Sequencing of Putative WGA Product from Hair Shaft Extracts

In an attempt to determine the accuracy of WGA methods, a subset of WGA product and multiplex PCR product from this WGA product were sequenced. However, there is no direct evidence that the resulting sequence data originates from the WGA product itself, since the proper controls were not run. The following sample types were sequenced:

1. WGA product from QIAGEN® REPLI-g® Mitochondrial DNA and Mini kit, no mtDNA-specific PCR amplification
2. WGA product from all four kits, multiplex amplified with both multiplexes
3. Extracts from all donors, no WGA, multiplex amplified with both multiplexes

A total of 53 samples were sequenced in this Illumina® MiSeq™ run, the quality metrics were as follows:

- 1046K clusters/mm^2 (higher than the recommended 800K but still acceptable)
- 91.5% of reads > Q30
- 87.71% of clusters passed filter (PF)

3.9.1 Whole Genome Amplified DNA Samples from Hair Shaft

WGA product from the QIAGEN® REPLI-g® Mitochondrial DNA and Mini kit that was not further PCR amplified was quantified and diluted according to two dilution

strategies. The first strategy was based on the mtDNA specific qPCR quantification values. In most cases, this resulted in no additional dilution prior to library preparation, except for the positive control of the Mitochondrial DNA kit. The second strategy was based on the Qubit® dsDNA quantifications: in these cases, the WGA product was diluted to 1 ng of dsDNA in each sample prior to library preparation.

Both dilution strategies of WGA product showed low NGS read coverage when mapped to the mtGenome. An average of 247,000 clusters were associated with these samples, and only a small percentage (0.004 - 2.029%) of clusters aligned to the mtGenome. This result was not unexpected, due to the differences in quantification values for the mtDNA-specific qPCR, and the non-specific Qubit® quantifications, combined with the fact that these samples were not further PCR amplified.

Although diluting the WGA product to 1 ng of dsDNA increases the number of clusters passing filter for almost all of these samples, it does not increase the percentage of clusters that align to the rCRS. An exception was the HL60 positive control for the Mitochondrial DNA kit, which showed an average coverage of approximately 1,000 reads in the qPCR dilution and 4,000 reads in the Qubit® dilution. Approximately 38 - 45% of the clusters associated with these samples aligned to the mtGenome. These data were compared to the sequences obtained with LPCR amplification and MiSeq™ sequencing (Appendix II). Some sequence differences were noted. These are positions that do not share a common base between two treatments, in this case WGA versus LPCR amplification of the same donor. The qPCR dilution showed one mixed position of 2.18% and the Qubit® dilution showed three mixed positions at frequencies between 1.22 and 1.57%. However, as stated previously, there is no evidence that these sequence data can

be contributed to the WGA process.

Potentially, these two WGA kits create primer hyperbranches in cases of insufficient DNA input into the WGA reaction, as explained in section 1.5.2. This non-specific product may then be tagmented by the library preparation and subsequently bridge amplified and sequenced by the NGS system, but does not generate mtDNA sequence data. This interpretation is supported by the observation that negative controls result in on average 1800 clusters with 0.27 - 3.55% of clusters aligning to the mtGenome, and the reagent blanks generate an average of 180,000 clusters with 0.01 - 0.61% of clusters aligning.

A sterile water sample was used as input for the Illumina® Nextera® XT library preparation method. Naturally, no clusters are expected for this sample. However, 1360 clusters passed filter for this sample, of which 7.6% aligned to the mtGenome, with a maximum read depth of 33. The majority of these reads align to areas targeted by the multiplex PCR primers. Considering 16 multiplex amplified samples were processed on the same 96-well plate with Nextera® XT, this indicates that the library preparation method may be sensitive to contamination if not handled with extreme care.

3.9.2 Multiplexed Whole Genome Amplified Samples

In contrast to the unamplified samples described above, high-coverage NGS data was generated for the multiplex amplified samples, with an average of 441,000 clusters passing filter, of which approximately 93% mapped back to the rCRS. The median coverage across the mtGenome was 7,151 reads for these samples, since median coverage

is calculated as the total number of reads over the entire mtGenome. However, coverage was up to 250,000 - 300,000 reads in regions covered by the multiplexes (Figure 24).

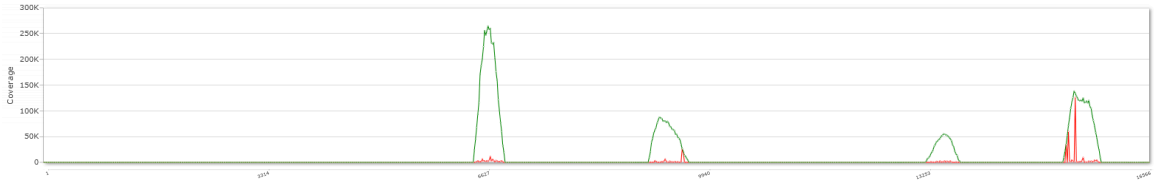


Figure 24: Coverage graph of WGA product amplified with multiplex PCR. HL60 processed with the Sigma-Aldrich® GenomePlex® WGA2 kit, diluted 1000x and amplified with MP5 primers. Trace derived from Illumina® MiSeq™ Reporter.

The accuracy of the sequence data derived from these multiplexed samples was determined by analysis in CGW (Table 25). Sequence differences arising from a comparison with LPCR data were noted as described in section 3.9.1. The PhiX sequencing control showed the expected sequence data. Thus, the direct comparison of NGS data from WGA material to LPCR product allowed for the observation of potentially WGA-induced effects on the DNA sequences. However, it cannot be excluded that differences arise from base misincorporations introduced by the tagmentation or PCR amplification steps in the library preparation method, and therefore differences cannot confidently be attributed to the WGA process.

Any sequence data obtained inside the multiplex primer binding regions or outside of the amplicon regions was discarded. Indel misalignments and small indels were not further analyzed. In addition, any low-level mixed positions below 1.0% were excluded from analysis, as it is yet unknown whether low-level mixtures below this threshold can be confidently called.

Table 25: Sequence differences in multiplex PCR data from WGA results. The NGS sequences from multiplexed WGA product were compared to the NGS sequences from LPCR product from the same donors, and differences or mixed positions between the corresponding mtDNA sequences were noted. Data derived from CGW. Q®R®: QIAGEN® REPLI-g®. S-A®: Sigma-Aldrich®. MP: multiplex primer set used for amplification. Zymo: purified with Zymo® Clean & Concentrator-5™ kit. Dil: diluted.

Template	Donor	MP	Total # positions	Highest frequency (%)	Notes
Hair shaft extract	002	5	0	0	-
	009	5	17	2.02	-
	020	5	0	0	-
	020	1	2	5.44	Possibly biological?
Q®R® Mitochondrial DNA kit	002-2	5	3	1.5	-
	020-2	5	6	1.61	-
Q®R® Mini kit	020-2	1	6	30.7	Contaminated?
	HL60	1	21	14.46	-
Q®R® Single Cell kit	020-1 Zymo	1	18	13	-
	020-1 100x Dil	1	18	49.69	-
	020-1 1000x Dil	1	17	52.87	-
	HL60 1000x Dil	1	15	9.16	-
S-A® GenomePlex® WGA2 kit	020 Zymo	5	44	5.85	-
	020 100x Dil	5	76	8.56	-
	020 1000x Dil	5	36	21.79	-
	HL60 1000x Dil	5	50	3.2	-

One of the Mini kit samples (020-2, MP1) may have been contaminated since the negative control from this WGA reaction showed many mtDNA copies upon qPCR quantification. In addition, multiple mixed positions are called at positions of known variation from the rCRS for this donor, which is consistent with a DNA mixture of more than one donor.

The hair shaft extract from donor 020, amplified with multiplex 1 primers, shows a mixed position at a frequency of 5.44%, which is below the detection threshold of 10% for Sanger sequencing. This likely did not occur due to base misincorporations as a result of the incorporation of an amplification primer that contains mismatches compared to the

template, but is still able to amplify this template. Nor did other low-level mixed positions accompany this position, as would be the case with contamination. MtDNA sequences have been shown to fluctuate throughout tissue types in an individual (Wilson et al. 1997). Since this particular hair extract was sequenced only once, it is possible that the observed 5.44% mixed position is of biological nature. However, this cannot be confirmed until further sequencing is performed.

It should be noted that the 100-fold and 1000-fold dilutions of Single Cell kit product derived from hair shaft extracts show mixed positions at higher percentages than is exhibited by neat or diluted product from the other three kits. In contrast, the GenomePlex® kit shows a higher number of mixed positions than any of the other three kits. However, as stated previously, there is no evidence that WGA material is responsible for any of these NGS results. Since the proper controls were not run, the possibility remains that the extracted hair shaft DNA provided the DNA template for these multiplex PCR amplifications.

CHAPTER 4: DISCUSSION

4.1 Sequence Comparison of Next Generation Sequencing vs. Sanger

In some instances in forensic casework nDNA may not be present in sufficient quantity or quality for STR analysis. In these cases, mtDNA is an excellent alternative. However, it is often too labor-intensive or expensive to sequence the entire mtGenome with traditional Sanger sequencing methods, and therefore analysis is frequently limited to the hypervariable regions. NGS allows for more cost-effective analysis of entire mtGenomes by obtaining more data points from the same sample. Here, a method is presented for the NGS analysis of entire mtGenomes from reference samples.

A long PCR approach successfully amplified the entire mitochondrial genome from buccal swabs from all eight donors. A single buccal swab generated ample amounts of DNA for LPCR processing: by using 200,000 copies of mtDNA from an extract as input, an average of 6 ng/µl of LPCR product was generated - Illumina® Nextera® XT requires only 1 ng of input.

However, since LPCR failed to amplify with DNA from cotton buccal swabs that had been stored at room temperature, but amplified well with fresh buccal swabs, it is possible that ongoing microbial activity allowed DNA degradation to occur. Thus, DNA extraction should be performed on buccal swabs that were freshly taken from the donor or stored on swabs with internal antimicrobial activity. Alternatively, extraction could be performed on cells that have been sloughed off from buccal swabs onto FTA cards, which exhibit antimicrobial activity and can be stored at room temperature for extended

periods of time (Whatman).

When compared to Sanger sequencing data, NGS data derived from long PCR amplicons generates the expected sequence. This includes known low-level mixed positions, for instance an 8% C to T transition in donor 001 at position 16,093, which is a known high mutation rate site (Tully et al. 2000)(Appendix I/II).

NGS data derived from the LPCR product span the entire mitochondrial genome at high coverage. The average coverage for all donor sequences was approximately 13,000x across the mtGenome. Due to the overlapping nature of the PCR primer sets, double sequence coverage was expected for the areas between nucleotides 15195 - 1892, which includes the non-coding region, as well as for 9397 - 9777. Elevated coverage was indeed seen in these regions, although coverage may be artificially lowered when using a non-circularized genome for mapping. An increase in coverage in these regions facilitates even deeper observation of low-level mixtures.

It should be noted that indels currently present a limitation in NGS analysis, as read mapping algorithms are known to misalign reads containing these indels. Mapping algorithms will likely be improved in the near future. Although small indels were ignored in analyses in this study, it was noted that a low-level mixed base at position 12,417, which is located in a region with eight adenines, is seen consistently at a frequency of approximately 4% in the NGS data from all donors, and across different runs. This may indicate that this homopolymer is causing a sequencing issue with the Illumina® MiSeq™ instrument or an alignment issue with these software packages. Although homopolymer regions are known to be heteroplasmic (Bendall and Sykes 1995), more research is needed to determine the origin of this mixed position.

As expected, the resolution with NGS is significantly higher compared to Sanger sequencing. Low-level sequence variation that had gone undetected in Sanger data was easily observed in the NGS data. In conclusion, the LPCR method is robust, easy to perform, and generates high coverage and high quality sequencing data. This makes it an excellent tool for generating whole mtGenome NGS sequence data from robust reference samples such as buccal swabs.

4.2 Evaluation of Illumina® Nextera® XT Library Preparation for Forensic Casework

Previously, library preparation presented one of the main bottlenecks in NGS analysis. Here, Illumina® Nextera® XT has shown to be an effective and rapid way of preparing libraries from different sources of DNA material for sequencing on the Illumina® MiSeq™ instrument, as libraries can prepared more quickly than with traditional library preparation methods. By incorporating indices, many samples can be run on the same flow cell, depending on the number of total indices available. The resulting sequence data can be demultiplexed, meaning that it is separated by its index reads and placed into distinct bins for each sample. The reported sequence data for each sample was as expected - exceptions to this are thought to arise mainly due to alignment issues.

As a measure of contamination in the WGA sequencing run a sterile water sample was prepared for sequencing alongside the WGA product and multiplex PCR products, and received a separate index. Since contamination is a concern that is always present in forensic casework, and even more so in instances where mtDNA is the focus of analysis,

this may be an appropriate control for the detection of contamination in an NGS run, which is extremely sensitive to contamination due to the high clonal amplification and deep sequencing that is implemented in NGS.

Upon sequencing, this control showed distinct contamination with multiplex PCR product, although three lanes on a 96-well plate separated the two sample types. This indicates that the control was contaminated with PCR product during the initial steps for tagmentation. This signal was likely increased during the suppression PCR in the Illumina® Nextera® XT protocol, as well as during bridge amplification during sequencing.

The coverage level of contamination that was seen in this control was as high as 35 reads in some areas of the mtGenome. As high-coverage data will likely be a requirement in forensic casework this level of contamination may be negligible. However, it may present a very real concern in low-coverage runs, and could influence data interpretation. Care should be taken with this library preparation method as to not cross-contaminate any materials that will be sequenced in the same run.

4.3 Evaluation of NIST Standards as Sequencing Controls

Two sets of NIST standards were sequenced without any prior amplification to evaluate their utility as controls to assess sequencing quality on a run-to-run basis. SRM 2392 is comprised of DNA extracts of the cell lines 9947A and CHR as well as the cloned HV1 region from CHR (Levin, Cheng, and Reeder 1999). SRM 2394 is comprised of two 285 bp amplicon amplified from CHR and 9947A, which differ by a single base pair at the same nucleotide position, and have been mixed at ten defined ratios ranging

from 1 to a 100% (Hancock, Tully, and Levin 2005).

The three SRM 2392 standards need to be amplified with mtDNA specific primers prior to sequencing. The extracts showed low coverage; nuclear DNA was present in these samples, which was also tagmented and sequenced. This resulted in a lower number of clusters mapping to mtDNA. The cloned HV1 region showed a similar problem: since the entire vector was sequenced, only a small portion of reads mapped back to the mtGenome. If these standards are to be used in the future, it is recommended that all samples be amplified before sequencing, whether with long PCR for the extracts, or HV1-specific primers for the cloned HV1 region. This will increase coverage, and thus increase resolution for these samples.

The SRM 2394 standards showed high sequence coverage for all mixtures. In addition, the observed frequencies were consistent with the expected frequencies reported by NIST. This indicates that these mixture standards may be implemented in future Illumina® MiSeq™ runs, to function as a sequencing control.

4.4 Whole Genome Amplification

Although mtDNA in reference samples is often robust, this may not be the case for challenging sample types such as hair shafts. Template mtDNA in these samples may be degraded, which does not support amplification with long PCR prior to NGS. Therefore, a different preparation of mtDNA in these samples may be necessary to obtain whole mtGenome NGS data. Here, mtDNA was pre-amplified with four different Whole Genome Amplification kits. The resulting product, as well as PCR product from a multiplex PCR amplification, was analyzed with NGS. The following conclusions are

based on limited data, and should therefore be considered as preliminary. Additional work needs to be performed in order to confirm these findings.

In these experiments, the evaluated WGA kits amplify mtDNA from hair extracts less efficiently than from buccal swabs, with the exception of the PCR-based Sigma-Aldrich® GenomePlex® WGA2 kit. Hair shaft extracts may contain degraded DNA, which has been stated to possibly decrease DNA yield in MDA reactions (Lage et al. 2003). As with buccal extracts, all kits except the GenomePlex® kit were inconsistent in amplification when comparing replicate samples from the same hair shaft extract, as well as replicate experiments using different extracts.

Hair shaft DNA extracts processed with the QIAGEN® REPLI-g® Mini and Mitochondrial DNA kits showed low sequence coverage upon sequencing with the Illumina® MiSeq™, however, due to the low mtDNA concentrations in these samples, these were not expected to generate high-coverage data. In contrast, a positive control prepared with the Mitochondrial DNA kit resulted in high coverage throughout the entire mtGenome. These results indicate that WGA product of these two kits, using hair shaft extracts as DNA template, cannot directly be sequenced with the two dilution strategies that were used for library preparation. Further studies need to be performed to optimize the input of WGA product into the NGS library preparation method.

These samples from the Mini and Mitochondrial DNA kit were relatively low in mtDNA concentration, yet the quantification of total dsDNA and ssDNA was significantly higher. It is possible that the random hexamers in these kits branch off each other due to low input in WGA, which contributes to the total dsDNA concentration in the sample (Lage et al. 2003). This could explain why only a small percentage of clusters

that have been assigned to each index actually align to the mtGenome. Potentially, a molecular crowding agent can be used to amplify these sample types more efficiently (Ballantyne et al. 2006).

Another solution for the low coverage in these samples could be evaluated by performing purification steps prior to sequencing. It is possible that by purifying these samples, primers and other reagents in the WGA product could be removed. In addition, the DNA template may be concentrated in a smaller volume. As such, the input of mtDNA into Illumina® Nextera® XT may be increased.

Amplification bias may be a concern with WGA techniques. Although pooling strategies were attempted to ameliorate the effects of possible unequal amplification, the coverage in these samples was too low to conclude whether any such amplification bias was present in these samples or whether the pooling strategy was an appropriate method of reducing bias.

It should be noted that due to their higher mtDNA concentrations, material from the Sigma-Aldrich® GenomePlex® WGA2 kit and the QIAGEN® REPLI-g® Single Cell kit may generate sequence data of a sufficient coverage for low-level mixture detection. However, material from these kits was not sequenced, due to concerns about their concentration with respect to other samples on the flow cell, as well as the possibility of cross-contamination due to frequent handling of these samples.

Long PCR amplification on hair extracts is not possible, as was expected due to the fragmented nature of the extracted mtDNA (Berger and Parson 2009). Previous, contradictory, studies have stated that certain WGA methods can facilitate the PCR amplification of larger targets than the fragment size of the starting material

(Maciejewska, Jakubowska, and Pawłowski 2013, Ballantyne, van Oorschot, and Mitchell 2007). In this study, a 9.1 kb long PCR amplification on WGA material of the QIAGEN® REPLI-g® Single Cell kit or the Sigma-Aldrich® GenomePlex® WGA2 kit was not successful.

However, multiplex amplification of relatively short, 400-700 bp mtDNA targets was successful on hair shaft extracts as previously shown by E. Burnside (Burnside *et al*, 2013). In forensic casework, sample material is often completely consumed in a single extraction process, and these DNA extracts may be of low mtDNA concentration. Although the hair shaft extractions performed in this study were often of sufficient concentration to support multiplex amplifications of the entire mtGenome, in some cases additional template may be necessary to support downstream processes.

Preliminary data from this research effort shows that the WGA kits that were evaluated all facilitate an augmentation in mtDNA copy number, potentially creating more template for subsequent analysis, albeit that some kits may be more efficient in pre-amplifying mtDNA than others.

Hair shaft extracts that were multiplex amplified provided the expected sequence using NGS as compared to the mtDNA sequence from these donors that was derived using Sanger sequencing, with the exception of one 5.44% mixed position that presents as a possibly biological variant in one of the samples. It is unlikely that this mixed position is a result of the incorporation of an amplification primer that contains base mismatches compared to the template. Nor was this position accompanied by other low-level mixed positions, as is expected with contamination. MtDNA sequences have been observed to differ throughout tissue types in an individual (Wilson et al. 1997). Since the

observed frequency is lower than the 10% detection threshold with Sanger sequencing, and this particular hair extract was sequenced with NGS only once, it is possible that the observed 5.44% mixed position is of biological nature. However, this cannot be confirmed until further sequencing is performed.

Error rates that result from the sequencing chemistry are low, as was evaluated by sequence data from the PhiX control. From these combined findings, it seems unlikely that mixed positions that were observed in the multiplexed PCR data originated during multiplexed PCR amplification or sequencing. However, as the PhiX control does not undergo library preparation, it cannot be excluded that the observed mixed positions are a result of base misincorporations introduced by the tagmentation or PCR amplification steps in the Illumina® Nextera® XT library preparation method. In addition, it cannot be confirmed that WGA product provided the template for these multiplexed PCR reactions, since the proper controls were not sequenced. Therefore, any mixed positions cannot confidently be attributed to the WGA process, and further studies are necessary to confirm these results.

The WGA product from the QIAGEN® REPLI-g® Mitochondrial DNA kit gives rise to a small number of low-percentage mixed positions. This could indicate that this kit gives the most accurate results, however it is important to note that these observations are based on limited data. As stated previously, it cannot be confirmed that WGA product provided the template for these sequencing reactions. Therefore, more studies are necessary to determine whether this result holds true for additional sequencing runs.

4.5 Concluding Remarks

To conclude, this research effort presents a method for obtaining whole mtGenome NGS data from reference samples, such as buccal swabs. A long PCR amplification was performed, after which PCR product was prepared with Illumina® Nextera® XT and sequenced on the Illumina® MiSeq™. This approach generated high-quality, high-coverage data, which reflected the data obtained by Sanger sequencing.

In addition, preliminary data was gathered as a basis for future work targeted towards NGS analysis of challenging sample types. Future work in this area needs to be performed. More replicate experiments with all four WGA kits should be performed to observe their augmentation of mtDNA from hair shaft extracts. In addition, WGA product from the QIAGEN® REPLI-g® Single Cell kit and the Sigma-Aldrich® GenomePlex® WGA2 kit should be sequenced without additional multiplex PCR amplification, to observe whether PCR amplification is necessary for obtaining high-coverage NGS data. Furthermore, if an NGS run is repeated with the right controls, a more detailed assessment can be made as to how accurately all four WGA methods amplify template mtDNA.

WORKS CITED

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11:R119.

Affymetrix®. 2011. USB® ExoSAP-IT® PCR Product Cleanup - Brief Protocol. Available from: http://media.affymetrix.com/support/technical/usb/brief_proto/78200B.pdf

Agencourt®. Agencourt® CleanSEQ® Dye-Terminator Removal. Available from: https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol%20000600v032.pdf

Agencourt®. Agencourt® AMPure® XP. Available from: https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol_000387v001.pdf

Agilent Technologies. 2006a. Agilent DNA 1000 Kit Guide. Available from: http://www.chem.agilent.com/Library/usermanuals/Public/G2938-90014_KitGuideDNA1000Assay_ebook.pdf

Agilent Technologies. 2006b. Agilent DNA 7500 and DNA 12000 Kit Guide. Available from: http://www.uri.edu/research/gsc/docs/DNA7500_Guide.pdf

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: Accurate indel calls from short-read data. Genome Res. 21:961–973.

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature 290:457.

Andréasson H, Nilsson M, Styrman H, Pettersson U, Allen M. 2007. Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. Forensic Sci. Int. Genet. 1:35–43.

Applied Biosystems®. 2006. VariantSEQr™ and mitoSEQr™ Resequencing Systems - Protocol. Available from: http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_041392.pdf

Applied Biosystems®. 2010. BigDye® Terminator v1.1 Cycle Sequencing Kit - Protocol. Available from: http://tools.invitrogen.com/content/sfs/manuals/cms_041330.pdf

Applied Biosystems®. 2012. PrepFiler® and PrepFiler® BTA Forensic DNA Extraction Kits. Available from: http://www3.appliedbiosystems.com/cms/groups/applied_markets_support/documents/generaldocuments/cms_096102.pdf

Applied Biosystems®. Quantifiler Kits: Quantifiler Human DNA Quantification Kit and Quantifiler Y Human Male DNA Quantification Kit - User's Manual. Available from: http://tools.invitrogen.com/content/sfs/manuals/cms_041395.pdf

Ballantyne KN, van Oorschot RAH, John Mitchell R, Koukoulas I. 2006. Molecular crowding increases the amplification success of multiple displacement amplification and short tandem repeat genotyping. Anal. Biochem. 355:298–303.

Ballantyne KN, van Oorschot RAH, Mitchell RJ. 2007. Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. Forensic Sci. Int. 166:35–41.

Barber AL, Foran DR. 2006. The Utility of Whole Genome Amplification for Typing Compromised Forensic Samples. J. Forensic Sci. 51:1344–1349.

Barker DL, Hansen MST, Faruqi AF, Giannola D, Irsula OR, Lasken RS, Latterich M, Makarov V, Oliphant A, Pinter JH, et al. 2004. Two Methods of Whole-Genome Amplification Enable Accurate Genotyping Across a 2320-SNP Linkage Panel. Genome Res. 14:901–907.

Barnes WM. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. 91:2216–2220.

Bendall KE, Sykes BC. 1995. Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am. J. Hum. Genet. 57:248–256.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Berger C, Parson W. 2009. Mini-midi-mito: Adapting the amplification and sequencing strategy of mtDNA to the degradation state of crime scene samples. Forensic Sci. Int. Genet. 3:149–153.

Blanchard JL, Schmidt GW. 1996. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. Mol. Biol. Evol. 13:537–548.

Blanco L, Bernad A, Lázaro JM, Martín G, Garmendia C, Salas M. 1989. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. J. Biol. Chem. 264:8935–8940.

Bogenhagen D, Clayton DA. 1974. The Number of Mitochondrial Deoxyribonucleic Acid Genomes in Mouse L and Human HeLa Cells: Quantitative Isolation of Mitochondrial Deoxyribonucleic Acid. J. Biol. Chem. 249:7991–7995.

Budowle B, Allard MW, Wilson MR, Chakraborty R. 2003. Forensics and mitochondrial DNA: applications, debates, and foundations. Annu. Rev. Genomics Hum. Genet. 4:119–141.

Budowle, B, Wilson, MR, DiZinno, JA. 1999. Interpretation guidelines for mitochondrial DNA sequencing. Proc. Tenth Int. Symp. Hum. Identif.

Burnside, E., Bintz, B., Wilson, M. 2012. An Optimized Protocol for Extraction of Mitochondrial DNA from Hair Shafts. Poster Present. Int. Symp. Hum. Identif. Nashv. TN.

Burnside, E., Bintz, B., Wilson, M. 2013. Next Generation Sequencing of the Human Mitochondrial Genome Using a Multiplexed PCR Strategy and Illumina Nextera XT. Poster Present. Int. Symp. Hum. Identif. Atlanta GA.

Butler JM. 2005. Forensic DNA Typing, Second Edition: Biology, Technology, and Genetics of STR Markers. 2nd ed. Academic Press.

Campbell NA, Reece JB. 2004. Biology. 7th ed. New York: Benjamin Cummings.

CLC bio. 2012. Read Mapping - White Paper. Available from: http://www.clcbio.com/files/whitepapers/whitepaper-on-CLC-read-mapper.pdf

CLC bio. CLC Manuals - Quality trimming. Available from: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Quality_trimming.html

Coble M, Just R, O'Callaghan J, Letmanyi I, Peterson C, Irwin J, Parsons T. 2004. Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int. J. Legal Med. 118:137–146.

Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38:1767–1771.

Collins PJ, Hennessy LK, Leibelt CS, Roby RK, Reeder DJ, Foxall PA. 2004. Developmental validation of a single-tube amplification of the 13 CODIS STR loci, D2S1338, D19S433, and amelogenin: the AmpFlSTR Identifiler PCR Amplification Kit. J. Forensic Sci. 49:1265–1277.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Davies PA, Gray G. 2002. Long-Range PCR. In: Theophilus BDM, Rapley R, editors. PCR Mutation Detection Protocols. Vol. 187. Methods in Molecular Biology. Humana Press. p. 51–55.

Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, et al. 2002. Comprehensive Human Genome Amplification Using Multiple Displacement Amplification. Proc. Natl. Acad. Sci. 99:5261–5266.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 23:4742–4743.

Fendt L, Zimmermann B, Daniaux M, Parson W. 2009. Sequencing strategy for the whole mitochondrial genome resulting in high quality sequences. BMC Genomics 10:139–139.

Foran DR. 2006. Relative degradation of nuclear and mitochondrial DNA: an experimental approach. J. Forensic Sci. 51:766–770.

Glenn T. 2011. Field guide to next-generation DNA sequencers, Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. Mol. Ecol. Resour. 11, 11:759, 759–769, 769.

Goto K, Nishino I, Hayashi YK. 2006. Rapid and accurate diagnosis of facioscapulohumeral muscular dystrophy. Neuromuscul. Disord. 16:256–261.

Greenberg BD, Newbold JE, Sugino A. 1983. Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene 21:33–49.

Gunnarsdóttir ED, Li M, Bauchet M, Finstermeier K, Stoneking M. 2011. High-throughput sequencing of complete human mtDNA genomes from the Philippines. Genome Res. 21:1–11.

Hancock DK, Tully LA, Levin BC. 2005. A Standard Reference Material to determine the sensitivity of techniques for detecting low-frequency mutations, SNPs, and heteroplasmies in mitochondrial DNA. Genomics 86:446–461.

Heid CA, Stevens J, Livak KJ, Williams PM. 1996. Real time quantitative PCR. Genome Res. 6:986–994.

Holland MM, Parsons TJ. 1999. Mitochondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Sci. Rev. 11.

Homer N, Merriman B, Nelson SF. 2009. BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLoS ONE 4:e7767.

Honeycutt R, Sobral BWS, McClelland M. 1997. Polymerase Chain Reaction (PCR) Detection and Quantification Using a Short PCR Product and a Synthetic Internal Positive Control. Anal. Biochem. 248:303–306.

Huang MM, Arnheim N, Goodman MF. 1992. Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. Nucleic Acids Res. 20:4567–4573.

Illumina®. 2010. Go where the biology takes you - Genome Analyzer IIx. Available from: http://res.illumina.com/documents/products/brochures/brochure_genome_analyzer.pdf

Illumina®. 2011. cBot: Fully automated clonal cluster generation for Illumina sequencing. Available from: http://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf

Illumina®. 2012. Nextera® XT DNA Sample Preparation Guide.

Illumina®. 2013a. MiSeq Reporter User Guide. Available from: http://support.illumina.com/documents/documentation/Software_Documentation/MiSeqReporter/MiSeqReporter_UserGuide_15028784_J.pdf

Illumina®. 2013b. Preparing DNA Libraries for Sequencing on the MiSeq®. Available from: http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq_preparingdnaformiseq_15039740_b.pdf

invitrogen. 2010. Qubit™ dsDNA HS Assay Kits. Available from: http://probes.invitrogen.com/media/pis/mp32851.pdf

invitrogen. 2011. Qubit® ssDNA Assay Kit. Available from: http://probes.invitrogen.com/media/pis/mp10212.pdf

Irwin JA, Parson W, Coble MD, Just RS. 2011. mtGenome reference population databases and the future of forensic mtDNA analysis. Forensic Sci. Int. Genet. 5:222–225.

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD. 2007. Extending assembly of short DNA sequences to handle error. Bioinforma. Oxf. Engl. 23:2942–2944.

Jeffreys AJ, Wilson V, Thein SL. 1985. Hypervariable “minisatellite” regions in human DNA. Nature 314:67–73.

Jünemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, et al. 2013. Updating benchtop sequencing performance comparison. Nat. Biotechnol. 31:294–296.

Kao W-C, Chan AH, Song YS. 2011. ECHO: A reference-free short-read error correction algorithm. Genome Res. 21:1181–1192.

Kavlick MF, Lawrence HS, Merritt RT, Fisher C, Isenberg A, Robertson JM, Budowle B. 2011. Quantification of Human Mitochondrial DNA Using Synthesized DNA Standards*. J. Forensic Sci. 56:1457–1463.

Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M. 1993. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. PCR Methods Appl. 3:13–22.

Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biol. 10:R83–R83.

Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. 2011. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. PLoS ONE 6:e28240.

Kumar S, Dudley J. 2007. Bioinformatics software for biologists in the genomics era. Bioinformatics 23:1713–1717.

Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, Dillon D, Segraves R, Vossbrinck B, González A, Pinkel D, et al. 2003. Whole Genome Analysis of Genetic Alterations in Small DNA Samples Using Hyperbranched Strand Displacement Amplification and Array–CGH. Genome Res. 13:294–307.

Levin BC, Cheng H, Reeder DJ. 1999. A Human Mitochondrial DNA Standard Reference Material for Quality Control in Forensic Identification, Medical Diagnosis, and Mutation Detection. Genomics 55:135–146.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Life Technologies. Qubit® ssDNA Assay Kit - Invitrogen. Available from: http://www.lifetechnologies.com/order/catalog/product/Q10212

Life Technologies. Qubit® ssDNA Assay Kit. Available from: http://www.lifetechnologies.com/order/catalog/product/Q10212

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. 2012. Comparison of Next-Generation Sequencing Systems. BioMed Res. Int. [Internet] 2012. Available from: http://www.hindawi.com/journals/bmri/2012/251364/abs/

Liu Z (John). 2011. Next Generation Sequencing and Whole Genome Selection in Aquaculture. E-book: John Wiley and Sons.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. 2012. Performance comparison of benchtop high-throughput sequencing platforms. Nat. Biotechnol. 30:434–+.

Maciejewska A, Jakubowska J, Pawłowski R. 2013. Whole genome amplification of degraded and nondegraded DNA for forensic purposes. Int. J. Legal Med. 127:309–319.

Maragh S, Jakupciak JP, Wagner PD, Rom WN, Sidransky D, Srivastava S, O'Connell CD. 2008. Multiple strand displacement amplification of mitochondrial DNA from clinical samples. BMC Med. Genet. 9:7.

Mardis ER. 2008. Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet. 9:387–402.

Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE. 2011. Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput Sequencing Libraries from Nanogram Quantities of DNA. Appl. Environ. Microbiol.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Meacham F, Boffelli D, Dhahbi J, Martin DIK, Singer M, Pachter L. 2011. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics 12:451.

Metzker ML. 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. 11:31–46.

Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol. 12:R112–R112.

Pareek CS, Smoczynski R, Tretyn A. 2011. Sequencing technologies and genome sequencing. J. Appl. Genet. 52:413–435.

Parson W, Dür A. 2007. EMPOP—A forensic mtDNA database. Forensic Sci. Int. Genet. 1:88–92.

Parsons T, Muniec D, Sullivan K, Woodyatt N, AllistonGreiner R, Wilson M, Berry D, Holland K, Weedn V, Gill P, et al. 1997. A high observed substitution rate in the human mitochondrial DNA control region. Nat. Genet. 15:363–368.

QIAGEN®. 2010. QIAamp® DNA Investigator Handbook. Available from: http://www.qiagen.com/resources/Download.aspx?id={6CA68723-3E4E-4AA0-AD3B-AE42DAC49A05}&lang=en&ver=1

QIAGEN®. 2011a. REPLI-g® Mini/Midi Handbook. Available from: http://www.qiagen.com/resources/Download.aspx?id={843654E0-2CCB-474B-B4B8-8744453ED5CB}&lang=en&ver=1

QIAGEN®. 2011b. REPLI-g® Mitochondrial DNA Handbook. Available from: http://www.qiagen.com/resources/Download.aspx?id={062381AC-018E-4E33-8FB3-3AA788271A0C}&lang=en&ver=1

QIAGEN®. 2012a. REPLI-g® Single Cell Handbook. Available from: http://www.qiagen.com/resources/Download.aspx?id={38FACA1C-64B0-4281-AAB3-AA8324BBD181}&lang=en&ver=1

QIAGEN®. 2012b. QIAamp DNA Mini and Blood Mini Handbook. Available from: http://www.qiagen.com/resources/Download.aspx?id={67893A91-946F-49B5-8033-394FA5D752EA}&lang=en&ver=1

Salas A, Lareu MV, Carracedo A. 2001. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. Int. J. Legal Med. 114:186–190.

Sanger F, Nicklen S, Coulson AR. 1977. DNA Sequencing with Chain-Terminating Inhibitors. Proc. Natl. Acad. Sci. U. S. A. 74:5463–5467.

Schneider P., Balogh K, Naveran N, Bogus M, Bender K, Lareu M, Carracedo A. 2004. Whole genome amplification—the solution for a common problem in forensic casework? Int. Congr. Ser. 1261:24–26.

Seki M, Hayashida N, Shinozaki K. 1996. Amplification of Long Targets of Approximately 50 kb from Cloned Cosmid Inserts of Arabidopsis thaliana. DNA Res. 3:107–108.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26:1135–1145.

Sigma-Aldrich®. 2012. GenomePlex® Complete Whole Genome Amplification (WGA) Kit - Technical Bulletin. Available from: http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Bulletin/wga2bul.pdf

Sigma-Aldrich®. Whole Genome Amplification Advisor. Available from: http://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/Brochure/1/wga_advisor.pdf

Spits C, Caignec CL, Rycke MD, Haute LV, Steirteghem AV, Liebaers I, Sermon K. 2006. Whole-genome multiple displacement amplification from single cells. Nat. Protoc. 1:1965–1970.

Szabo S, Jaeger K, Fischer H, Tschachler E, Parson W, Eckhart L. 2012. In situ labeling of DNA reveals interindividual variation in nuclear DNA breakdown in hair and may be useful to predict success of forensic genotyping of hair. Int. J. Legal Med. 126:63–70.

TaKaRa Bio Inc. 2013. TaKaRa LA Taq®. Available from: http://www.clontech.com/takara/US/Products/PCR_Products/High_Performance_PCR/ibcGetAttachment.jsp?cItemId=10237&fileId=6632654&sitex=10031:22372:US

Tang S, Wang J, Zhang VW, Li F-Y, Landsverk M, Cui H, Truong CK, Wang G, Chen LC, Graham B, et al. 2013. Transition to Next Generation Analysis of the Whole Mitochondrial Genome: A Summary of Molecular Defects. Hum. Mutat. 34:882–893.

Taylor RW, Turnbull DM. 2005. Mitochondrial DNA mutations in human disease. Nat. Rev. Genet. 6:389–402.

Thermo Scientific. phiX174 RF1 DNA. Available from: http://www.thermoscientificbio.com/molecular-cloning/phix174-rf1-dna/

Tully LA, Parsons TJ, Steighner RJ, Holland MM, Marino MA, Prenger VL. 2000. A Sensitive Denaturing Gradient-Gel Electrophoresis Assay Reveals a High Frequency of Heteroplasmy in Hypervariable Region 1 of the Human mtDNA Control Region. Am. J. Hum. Genet. 67:432–443.

Voelkerding KV, Dames S, Durtschi JD. 2010. Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy. J. Mol. Diagn. 12:539–551.

De Vries A, Zwaan C, Beverloo H, Wagner A, Lankester A, te Boekhorst P, de Coo I, Schoonderwoerd G, Hellebrekers D, Hendrickx A. 2012. Mitochondrial DNA Mutations Are Involved in the Development of Childhood Myelodysplastic Syndrome. Mol. Determinants Juv. Myelomonocytic Leuk. Child. Myelodysplastic Syndr.:105.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available from: http://www.genome.gov/sequencingcosts/

Whatman. Whatman FTA Protocol BD01 - Applying and Preparing Blood Samples on FTA® Cards for DNA Analysis. Available from: http://www.whatman.com/UserFiles/File/Protocols/Bioscience/BD01%20-%20DNA%20-%20Applying%20and%20Preparing%20Blood%20Samples%20on%20FTA%20Cards.pdf

Wilson MR, Polanskey D, Butler J, DiZinno JA, Replogle J, Budowle B. 1995. Extraction, PCR amplification and sequencing of mitochondrial DNA from human hair shafts. BioTechniques 18:662–669.

Wilson MR, Polanskey D, Replogle J, DiZinno JA, Budowle B. 1997. A family exhibiting heteroplasmy in the human mitochondrial DNA control region reveals both somatic mosaicism and pronounced segregation of mitotypes. Hum. Genet. 100:167.

Yiping He, Jian Wu, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz Jr. LA, Kinzler KW, Vogelstein B, Papadopoulos N. 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464:610–614.

Zymo Research. DNA Clean & Concentrator™-5 Instruction Manual.

APPENDIX I: SANGER SEQUENCING REFERENCE DATA OF ALL DONORS

The below tables show the reference data that was obtained with Sanger sequencing for all donors in this study.

Table I.1: Sanger Sequence Donor 001

Position	**rCRS**	**Variant**
73	A	G
185	G	A
228	G	A
263	A	G
295	C	T
309.1	:	C
315.1	:	C
462	C	T
489	T	C
523	A	:
525	C	:
750	A	G
1,438	A	G
2,706	A	G
3,010	G	A
3,107	N	:
4,216	T	C
4,769	A	G
7,028	C	T
8,860	A	G
8,865	G	A
10,398	A	G
11,251	A	G
11,719	G	A
12,612	A	G
13,708	G	A
13,934	C	T
14,766	C	T
14,798	T	C
15,326	A	G

15,452	C	A
16,069	C	T
16,093	T	C/T
16,126	T	C
16,390	G	A
Total Differences		35

Table I.2: Sanger Sequence Donor 002

Position	**rCRS**	**Variant**
73	A	G
152	T	C
199	T	C
204	T	C
207	G	A
250	T	C
263	A	G
309.1	:	C
315.1	:	C
573.1	:	C
573.2	:	C
573.3	:	C
750	A	G
1,438	A	G
1,719	G	A
2,706	A	G
2,835	C	A
3,107	N	:
4,529	A	T
4,769	A	G
7,028	C	T
7,055	A	T
8,251	G	A
8,860	A	G
9,548	G	A
10,034	T	C
10,238	T	C
10,398	A	G
11,065	A	G
11,719	G	A

12,501	G	A
12,705	C	T
13,780	A	G
14,766	C	T
15,043	G	A
15,326	A	G
15,673	A	G/A
15,758	A	G
15,924	A	G
16,074	A	G
16,129	G	A
16,145	G	A
16,223	C	T
16,391	G	A
16,519	T	C
Total Differences		45

Table I.3: Sanger Sequence Donor 003

Position	**rCRS**	**Variant**
73	A	G
150	C	T
152	T	C
263	A	G
295	C	T
315.1	:	C
489	T	C
750	A	G
1,438	A	G
2,706	A	G
3,107	N	:
4,216	T	C
4,769	A	G
5,633	C	T
6,830	C	A
7,028	C	T
7,476	C	T
7,771	A	G
8,095	A	G
8,860	A	G
10,172	G	A

10,398	A	G
11,251	A	G
11,719	G	A
12,612	A	G
12,715	A	G
13,708	G	A
14,766	C	T
15,257	G	A
15,326	A	G
15,452	C	A
15,812	G	A
16,069	C	T
16,126	T	C
16,193	C	T
16,195	T	C
16,221	C	T
16,242	C	A
16,319	G	A
16,357	T	C
16,526	G	A
Total Differences		41

Table I.4: Sanger Sequence Donor 006

Position	**rCRS**	**Variant**
152	T	C
263	A	G
309.1	:	C
315.1	:	C
750	A	G
1,438	A	G
3,107	N	:
4,769	A	G
8,860	A	G
9,129	C	T
10,394	C	T
10,685	G	G/A
11,054	C	T
12,172	A	G
15,326	A	G
16,359	T	C

16,519	T	C
Total Differences		16

Table I.5: Sanger Sequence Donor 009

Position	**rCRS**	**Variant**
73	A	G
263	A	G
309.1	:	C
315.1	:	C
523	A	:
525	C	:
750	A	G
1,438	A	G
3,107	N	:
3,992	C	T
4,769	A	G
5,004	T	C
8,584	G	A
8,860	A	G
9,123	G	A
9,276	G	A
11,410	T	C
15,326	A	G
16,248	C	T
Total Differences		19

Table I.6: Sanger Sequence Donor 015

Position	**rCRS**	**Variant**
73	A	G
204	T	C
263	A	G
309.1	:	C
315.1	:	C
447	C	G
489	T	C
750	A	G
1,438	A	G
1,780	T	C
2,706	A	G

3,107	N	:
4,769	A	G
5,252	G	A
5,821	G	A
7,028	C	T
7,961	T	C
8,269	G	A
8,396	A	G
8,490	T	C
8,502	A	G
8,701	A	G
8,860	A	G
9,540	T	C
9,758	T	C
10,398	A	G
10,400	C	T
10,873	T	C
11,083	A	G
11,719	G	A
12,705	C	T
12,810	A	G
13,204	G	A
13,651	A	G
14,766	C	T
14,783	T	C
15,043	G	A
15,256	A	G
15,301	G	A
15,326	A	G
15,479	T	C
15,670	T	C
15,758	A	G
16,223	C	T
16,224	T	C
16,270	C	T
16,274	G	A
16,319	G	A
16,352	T	C
16,519	T	C
Total Differences		50

Table I.7: Sanger Sequence Donor 020

Position	**rCRS**	**Variant**
73	A	G
195	T	A
263	A	G
309.1	:	C
315.1	:	C
489	T	C
523	A	:
525	C	:
750	A	G
1,438	A	G
2,706	A	G
3,107	N	:
3,173	G	A
4,769	A	G
7,028	C	T
8,701	A	G
8,860	A	G
9,540	T	C
9,566	C	T
10,398	A	G
10,400	C	T
10,873	T	C
11,719	G	A
12,007	G	A
12,705	C	T
13,135	G	A
14,766	C	T
14,783	T	C
15,043	G	A
15,301	G	A
15,326	A	G
15,431	G	A
16,223	C	T
16,234	C	T
16,362	T	C
16,519	T	C
Total Differences		36

Table I.8: Sanger Sequence Donor 021

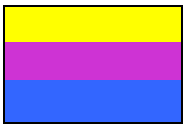
Position	rCRS	Variant
73	A	G
183	A	G
249	A	:
263	A	G
290	A	:
291	A	:
309.1	:	C
315.1	:	C
489	T	C
493	A	G
523	A	:
525	C	:
750	A	G
1,438	A	G
2,706	A	G
3,107	N	:
3,552	T	A
4,715	A	G
4,769	A	G
7,028	C	T
7,196	C	A
7,948	C	T
8,584	G	A
8,701	A	G
8,860	A	G
9,540	T	C
9,545	A	G
10,398	A	G
10,400	C	T
10,873	T	C
11,719	G	A
11,914	G	A
12,696	T	C
12,705	C	T
13,263	A	G
14,022	A	G
14,318	T	C
14,766	C	T
14,783	T	C

15,043	G	A
15,301	G	A
15,326	A	G
15,487	A	T
16,223	C	T
16,298	T	C
16,325	T	C
16,327	C	T
16,345	A	T
Total Differences		48

APPENDIX II: NGS DATA FROM ALL DONORS

The below tables show the Illumina® MiSeq™ data derived from long PCR amplification and NGS of all donors in this study. Data was analyzed with MiSeq™ Reporter 2.2 (MSR).

Table II.1: Legend for interpretation of NGS data tables.



Expected variant from the rCRS
Low-level mixed position
Low-level mixed position in homopolymer region

Table II.2: Illumina® MiSeq™ Data Donor 001

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	4091
185	SNP	G->GA	99	4876
228	SNP	G->GA	9	3678
263	SNP	A->AG	100	2422
295	SNP	C->CT	100	1965
302	Indel	-/C	64	1268
310	Indel	-/C	100	989
462	SNP	C->CT	100	3535
489	SNP	T->TC	100	3038
513	Indel	CA/--	76	2541
750	SNP	A->AG	100	6281
1438	SNP	A->AG	100	8236
2706	SNP	A->AG	100	2373
3010	SNP	G->GA	100	5558
3106	Indel	N/-	93	5672
4216	SNP	T->TC	100	8719
4769	SNP	A->AG	100	8988
7028	SNP	C->CT	99	6746
8860	SNP	A->AG	100	8540

8865	SNP	G->GA	100	8752
10398	SNP	A->AG	100	3272
11251	SNP	A->AG	100	4863
11719	SNP	G->GA	100	5384
12612	SNP	A->AG	100	3561
13708	SNP	G->GA	100	1600
13934	SNP	C->CT	100	5167
14766	SNP	C->CT	100	5021
14798	SNP	T->TC	100	5227
15326	SNP	A->AG	100	10169
15452	SNP	C->CA	100	11480
16069	SNP	C->CT	100	8825
16093	SNP	T->TC	92	10463
16126	SNP	T->TC	100	11615
16390	SNP	G->GA	100	11826

Table II.3: Illumina® MiSeq™ Data Donor 002

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	11289
152	SNP	T->TC	100	16694
199	SNP	T->TC	100	10632
204	SNP	T->TC	100	9558
207	SNP	G->GA	100	9341
250	SNP	T->TC	100	5959
263	SNP	A->AG	100	4512
302	Indel	-/C	91	1755
310	Indel	-/C	100	2043
567	Indel	---/CCC	49	1781
750	SNP	A->AG	100	18207
1438	SNP	A->AG	100	25567
1719	SNP	G->GA	100	24450
2706	SNP	A->AG	100	6461
2835	SNP	C->CA	100	11764
3106	Indel	N/-	94	10710
4529	SNP	A->AT	100	10163
4769	SNP	A->AG	100	11051
7028	SNP	C->CT	99	13846
7055	SNP	A->AT	100	12759
8251	SNP	G->GA	100	9854
8843	SNP	T->TC	2	16098
8860	SNP	A->AG	100	15077
9548	SNP	G->GA	100	9238
10034	SNP	T->TC	100	7260
10238	SNP	T->TC	100	6587

10398	SNP	A->AG	100	8828
11065	SNP	A->AG	100	11494
11719	SNP	G->GA	100	14204
12501	SNP	G->GA	100	8863
12705	SNP	C->CT	100	11234
13780	SNP	A->AG	100	4520
14766	SNP	C->CT	100	12300
15043	SNP	G->GA	100	16826
15326	SNP	A->AG	100	28723
15673	SNP	A->AG	83	26461
15758	SNP	A->AG	100	26543
15924	SNP	A->AG	100	20390
16074	SNP	A->AG	100	20066
16129	SNP	G->GA	99	23467
16145	SNP	G->GA	1	24327
16223	SNP	C->CT	99	31446
16391	SNP	G->GA	100	31781
16519	SNP	T->TC	100	11915

Table II.4: Illumina® MiSeq™ Data Donor 003

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	12352
150	SNP	C->CT	100	18709
152	SNP	T->TC	100	18589
263	SNP	A->AG	100	6161
295	SNP	C->CT	100	4621
310	Indel	-/C	100	2800
489	SNP	T->TC	100	6026
750	SNP	A->AG	100	19878
1438	SNP	A->AG	100	25184
2706	SNP	A->AG	100	6989
3106	Indel	N/-	94	11711
4216	SNP	T->TC	100	13823
4769	SNP	A->AG	100	12363
5633	SNP	C->CT	100	7407
6830	SNP	C->CA	100	16315
7028	SNP	C->CT	100	14728
7476	SNP	C->CT	100	7281
7771	SNP	A->AG	100	13124
8095	SNP	A->AG	100	11163
8860	SNP	A->AG	100	16142
10172	SNP	G->GA	100	5235
10398	SNP	A->AG	100	7142
11251	SNP	A->AG	100	11689

11719	SNP	G->GA	100	11450
12612	SNP	A->AG	100	9729
12715	SNP	A->AG	100	10944
13708	SNP	G->GA	100	3581
14766	SNP	C->CT	100	9727
15257	SNP	G->GA	100	20862
15326	SNP	A->AG	100	25206
15452	SNP	C->CA	100	24965
15812	SNP	G->GA	100	24231
16069	SNP	C->CT	100	21264
16126	SNP	T->TC	100	25162
16193	SNP	C->CT	100	31445
16195	SNP	T->TC	100	32680
16221	SNP	C->CT	99	30154
16242	SNP	C->CA	100	31311
16319	SNP	G->GA	99	23878
16357	SNP	T->TC	100	27890
16526	SNP	G->GA	100	9901

Table II.6: Illumina® MiSeq™ Data Donor 006

Position	Variant Type	Call	Frequency	Depth
152	SNP	T->TC	97	24389
263	SNP	A->AG	1	5809
302	Indel	-/C	54	2264
310	Indel	-/C	84	2365
750	SNP	A->AG	100	26627
1438	SNP	A->AG	100	34013
3106	Indel	N/-	94	15518
4769	SNP	A->AG	100	15148
8860	SNP	A->AG	100	18638
9129	SNP	C->CT	100	13532
10394	SNP	C->CT	100	9577
10685	SNP	G->GA	12	14805
11054	SNP	C->CT	100	12260
12172	SNP	A->AG	100	13435
15326	SNP	A->AG	100	32190
16359	SNP	T->TC	100	31132
16519	SNP	T->TC	100	14535

Table II.6: Illumina® MiSeq™ Data Donor 009

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	14785
215	SNP	A->AG	4	13661
263	SNP	A->AG	100	7286
302	Indel	-/C	90	2938
310	Indel	-/C	100	3204
513	Indel	CA/--	81	7516
750	SNP	A->AG	100	24436
1438	SNP	A->AG	100	29430
3106	Indel	N/-	94	13618
3992	SNP	C->CT	100	15232
4769	SNP	A->AG	100	14626
5004	SNP	T->TC	100	16469
8584	SNP	G->GA	100	11174
8860	SNP	A->AG	100	18337
9123	SNP	G->GA	100	22702
9276	SNP	G->GA	100	16198
11410	SNP	T->TC	100	17430
15326	SNP	A->AG	100	30673
16248	SNP	C->CT	100	34358

Table II.7: Illumina® MiSeq™ Data Donor 015

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	15397
204	SNP	T->TC	100	15078
263	SNP	A->AG	100	7112
302	Indel	-/C	91	2810
310	Indel	-/C	100	3052
447	SNP	C->CG	100	6649
489	SNP	T->TC	100	5794
750	SNP	A->AG	100	23475
921	SNP	T->TC	2	25183
1438	SNP	A->AG	100	30130
1780	SNP	T->TC	100	25359
2706	SNP	A->AG	100	7247
3106	Indel	N/-	95	12924
4769	SNP	A->AG	100	13278
5252	SNP	G->GA	100	13155
5821	SNP	G->GA	100	9269
7028	SNP	C->CT	99	15464
7961	SNP	T->TC	100	15979

8269	SNP	G->GA	100	11004
8396	SNP	A->AG	100	10300
8490	SNP	T->TC	100	10724
8502	SNP	A->AG	100	10297
8701	SNP	A->AG	100	16737
8860	SNP	A->AG	100	17238
9540	SNP	T->TC	100	13209
9758	SNP	T->TC	100	10955
10398	SNP	A->AG	100	8932
10400	SNP	C->CT	100	9012
10873	SNP	T->TC	100	7475
11083	SNP	A->AG	100	13349
11719	SNP	G->GA	100	15239
12705	SNP	C->CT	100	12410
12810	SNP	A->AG	100	11792
13204	SNP	G->GA	100	16214
13651	SNP	A->AG	100	4511
14766	SNP	C->CT	100	12807
14783	SNP	T->TC	100	13259
15043	SNP	G->GA	100	17550
15256	SNP	A->AG	100	25324
15301	SNP	G->GA	100	27441
15326	SNP	A->AG	100	31776
15479	SNP	T->TC	100	29744
15670	SNP	T->TC	100	31170
15758	SNP	A->AG	100	30079
16223	SNP	C->CT	100	32348
16224	SNP	T->TC	100	34657
16270	SNP	C->CT	100	33167
16274	SNP	G->GA	99	32488
16319	SNP	G->GA	100	27380
16352	SNP	T->TC	100	32237
16519	SNP	T->TC	100	13959

Table II.8: Illumina® MiSeq™ Data Donor 020

Position	Variant Type	Call	Frequency	Depth
73	SNP	A->AG	100	12299
195	SNP	T->TA	100	13177
263	SNP	A->AG	100	4882
302	Indel	-/C	82	2057
310	Indel	-/C	100	2069
489	SNP	T->TC	100	3800
513	Indel	CA/--	84	4132

750	SNP	A->AG	100	17697
1438	SNP	A->AG	100	23124
2706	SNP	A->AG	100	6809
3106	Indel	N/-	95	11981
3173	SNP	G->GA	100	10440
4769	SNP	A->AG	100	12677
6734	SNP	G->GA	2	15873
7028	SNP	C->CT	100	15055
8701	SNP	A->AG	100	15661
8860	SNP	A->AG	100	16479
9540	SNP	T->TC	100	10361
9566	SNP	C->CT	100	8561
10398	SNP	A->AG	100	5759
10400	SNP	C->CT	100	5839
10873	SNP	T->TC	100	4543
11198	SNP	A->AT	5	9035
11719	SNP	G->GA	100	10134
12007	SNP	G->GA	100	10171
12705	SNP	C->CT	100	7986
13135	SNP	G->GA	100	7264
14536	SNP	A->AT	5	7899
14766	SNP	C->CT	100	8504
14783	SNP	T->TC	100	8880
15043	SNP	G->GA	100	12055
15301	SNP	G->GA	100	20890
15326	SNP	A->AG	100	24217
15431	SNP	G->GA	100	20827
16223	SNP	C->CT	100	27425
16234	SNP	C->CT	100	28294
16362	SNP	T->TC	100	26053
16519	SNP	T->TC	100	10670

Table II.9: Illumina® MiSeq™ Data Donor 021

Position	Variant Type	Call	Frequency
73	SNP	A->AG	100
183	SNP	A->AG	100
247	Indel	A/-	88
263	SNP	A->AG	1
285	Indel	AA/--	76
302	Indel	-/C	100
310	Indel	-/C	100
489	SNP	T->TC	100
493	SNP	A->AG	100

513	Indel	CA/--	83
750	SNP	A->AG	100
1438	SNP	A->AG	100
2706	SNP	A->AG	100
3106	Indel	N/-	94
3421	SNP	G->GA	2
3552	SNP	T->TA	100
4715	SNP	A->AG	100
4769	SNP	A->AG	100
7028	SNP	C->CT	100
7196	SNP	C->CA	100
7948	SNP	C->CT	100
8584	SNP	G->GA	100
8701	SNP	A->AG	100
8860	SNP	A->AG	100
9540	SNP	T->TC	100
#9545	SNP	A->AG	100
10398	SNP	A->AG	100
10400	SNP	C->CT	100
10873	SNP	T->TC	100
11719	SNP	G->GA	100
11914	SNP	G->GA	100
12696	SNP	T->TC	100
12705	SNP	C->CT	100
13263	SNP	A->AG	100
14022	SNP	A->AG	100
14318	SNP	T->TC	100
14766	SNP	C->CT	100
14783	SNP	T->TC	100
15043	SNP	G->GA	100
15301	SNP	G->GA	100
15326	SNP	A->AG	100
15487	SNP	A->AT	100
16223	SNP	C->CT	100
16298	SNP	T->TC	100
16325	SNP	T->TC	100
16327	SNP	C->CT	100
16345	SNP	A->AT	100